Humboldt-Universität zu Berlin
Institut für Bibliotheks- und Informationswissenschaft

Dissertation

Drivers and Barriers in Digital Scholarly Communication

Zur Erlangung des akademischen Grades

Doktor philosophiae (Dr. phil.)

eingereicht an der

Philosophischen Fakultät I

von

Sünje Dallmeier-Tiessen

Dekan: Prof. Michael Seadle, PhD

Gutachter: 1. Prof. Dr. Peter Schirmbacher
2. Prof. Michael Seadle, PhD
3. Dr. Salvatore Mele

Datum der Einreichung: 08.10.2012
Datum der Verteidigung: 28.01.2013

Zusammenfassung

Das digitale Zeitalter hat die Möglichkeit wissenschaftliche Abläufe und die wissenschaftliche Kommunikation nachhaltig zu verändern. In Anbetracht der wachsenden Nachfrage auf gesellschaftlicher und politischer Ebene nach Open Science, untersucht diese Arbeit Anreize und Hindernisse beim offenen Austausch von Forschungsmaterialien. Dabei wird der gegenwärtige Stand analysiert und Zukunftsperspektiven aufgezeigt, um dann entsprechende unterstützende Maßnahmen zu diskutieren.

Zwei Innovationen innerhalb von Open Science werden fokussiert untersucht: Open Access und der Umgang mit Forschungsdaten. Die Ergebnisse zeigen ein sehr positives Meinungsbild gegenüber beiden Innovationen, was sich allerdings nicht in einer übergreifenden Umsetzung in der Wissenschaft niederschlägt. Die disziplinären Unterschiede sind markant. Nichtsdestotrotz lassen sich übergeordnet abstrakte Ebenen herausarbeiten: Soziologische, technische & infrastrukturelle, sowie strategische & monetäre Aspekte gehören hierzu, wobei starke Interdependenzen zu verorten sind. Traditionell werden Qualität und Prestige von veröffentlichten wissenschaftlichen Ergebnissen als Maßgabe für die Reputation eines Wissenschaftlers angesehen, was klar in den Resultaten dieser Arbeit reflektiert ist. Wissenschaftler präferieren die Nutzung von Publikationsorganen und Arbeitsabläufen, die in der Fachgemeinschaft akzeptiert sind. Daraus folgt ein zögerlicher Umgang mit neuen Innovationen, z.B. dem offenem Zugang zu Forschungsdaten, wo es nur wenige etablierte Abläufe gibt. In der Diskussion dieser Arbeit wird die Notwendigkeit einer Verbindung zu heutigen Anreizsystemen und damit den Evaluierungssystemen in der Wissenschaft herausgestellt. Neue Strategien diesbezüglich sind im Aufbau, z.B. mit “zählbaren” Publikationen und Zitationen für Forschungsdaten und angepassten Metriken.

Die Kernthemen der gesellschaftlichen Ebene wurden in der Fallstudie der Hochenergiephysik genauer untersucht. Die Verfügbarkeit einer digitalen Bibliothek erlaubte dort zudem die praktische Implementierung von Open Science Werkzeugen. Die Ergebnisse unterstreichen das Potential solcher Ansätze: mit gezielten Diensten und Anreizen können Wissenschaftler für Open Science gewonnen werden; in diesem Fall zur Teilnahme in einem Crowdsourcingprojekt der digitalen Bibliothek und zur Umsetzung von „data sharing“. Dem Informationsmanagement kommt dabei eine neue Rolle und engere Zusammenarbeit mit der Wissenschaft zu. Die enge Betreuung von Wissenschaftlern im digitalen Forschungsumfeld kann für die Serviceentwicklung und –begleitung genutzt werden.

Damit sind die Ergebnisse für die meisten Akteure im wissenschaftlichen Publikationsprozess interessant. Die Partizipation in Open Science kann mit Hilfe von dezidierten Arbeitsabläufen gefördert werden. Des Weiteren ist es nun wichtig, die Anreizmechanismen weiter auszubauen, insbesondere auf Seiten der Forschungsevaluation.

Schlagwörter: Open Access, Forschungsdaten, Informationsmanagement, Open Science, Digitale Wissenschaft

Abstract

This thesis frames a new and comprehensive picture of digital scholarly communication today. The advent and pervasiveness of the internet could profoundly impact the research workflows, paving a way for Open Science. Given the growing demand by funders, society and policy-makers, it is needed to understand if this is indeed the case. Related research has suggested that adoption of innovations is not widespread. Thus, a more solid evidence base was needed to understand the current status and investigate drivers and barriers, to enable a more targeted support for Open Science in the future.

Two major Open Science innovations, Open Access and research data sharing, have been studied in detail in this thesis. A large-scale survey and personal interviews are used to gain detailed insights from a range of disciplines. In addition, a case study in the High Energy Physics (HEP) community was used to study the results in practice.

The results show that a rather positive attitude towards both, Open Access and research data sharing is not reflected in the researchers' practices. Disciplinary differences prevail and relate to the different publishing cultures and research workflows. Furthermore, it is shown that quality and prestige of research output are perceived as very important in determining a researcher's reputation. Researchers prefer community-approved publication outlets. They hesitate to explore new innovations, such as data sharing, for which only few established workflows exist in digital scholarly communication. The overall results point to a complex framework consisting of layers of societal, technical infrastructural, funding and strategy elements which are strongly interdependent. Interviewees highlight the significance of a link between such approaches on the one hand and the current incentive system and the research assessment schemes on the other.

Moreover, the results indicate that barriers can be overcome. In the case study, a strong collaboration with the community facilitated enhanced feedback loops to develop tailored and targeted services for Open Science. Researchers in the case study were successfully engaged in new innovative workflows: a crowdsourcing tool and data sharing in a digital library.

The results are of wider significance for stakeholders in digital scholarly communication today: they highlight that opportunities of Open Science are not yet explored widely. But with targeted support, it is possible to build on best practices and develop strategies that engage communities in new innovations. The results furthermore demand new strategies to establish links from Open Science services to the academic incentive system. It is needed to revisit the current research assessment scheme in regard to potential support mechanisms for Open Science.

Keywords: Open Access, Research Data, Information Management, Open Science, Digital Scholarly Communication

Table of Contents

Acknowledgements

This thesis would not have been possible without the support by several people, groups and collaborations.

In particular, I would like to thank Professor Schirmbacher for taking over the supervision and his continuous input throughout the thesis project. Similarly, I would like to thank Salvatore Mele for his inspiring ideas and fruitful discussions which helped me defining and streamlining the content of this thesis.

Parts of this research benefited from my participation in large-scale projects. I would thus thank my wonderful colleagues from the SOAP and ODE-Project. In that regard, it is also needed to point to the INSPIRE collaboration: I am delighted to be part of such an extraordinary team that tackles more than just the “everyday” challenges of a digital library. I had a wonderful time with my office mates and I need to thank them, and in particular Patricia and Henning, for their patience, support and laughter during this time.

I need to highlight Heinz, Robert and Viola who helped significantly to proofread my thesis. I am very grateful for their input and detailed corrections. I would in particular thank Heinz, for continuous discussions over the past years which not only resulted in research discussions, but also in reflections on “nearly everything”. Thanks also to Maxi, Sandra and Paul who helped me throughout my time at IBI.

I am indebted to my family and friends for their continuous support and patience. Some of them did so remotely, and I need to underline that without the support and trust of my family this project would not have been possible – never mind the distance.

Special thanks needs to be given to my friends. It is a pleasure to say thank you - I am grateful to have such wonderful friends here in Geneva and abroad, let it be in Berlin, Bremen, Bristol, Hamburg, Lugano, Münster and many other places. And, of course, special thanks shall also be given to Cat Power.

List of Abbreviations

CERN	European Organization for Nuclear Research
CMS	Compact Muon Solenoid (Experiment on the LHC)
DESY	Deutsches Elektronen-Synchrotron (German Electron Synchrotron)
DFG	Deutsche Forschungsgemeinschaft (German Research Foundation)
DOAJ	Directory Open Access Journals
DOI	Digital Object Identifier
DPHEP	Study Group for Data Preservation and Long Term Analysis in High Energy Physics
HEP	High Energy Physics
HSS	Humanities and Social Sciences
JCR	Journal Citation Report
LHC	Large Hadron Collider
MARC	Machine Readable Cataloging
NSF	National Science Foundation
OA	Open Access
ODE	Opportunities in Data Exchange (FP7 Project)
OpenDOAR	Directory of Open Access Repositories
ORCID	Open Researcher and Contributor ID
Parse.Insight	Permanent Access to the Records of Science in Europe (FP7 Project)
PLOS	Public Library of Science
ROAR	Registry of Open Access Repositories
SLAC	SLAC National Accelerator Laboratory (Stanford Linear Accelerator Center)
SCOAP3	Sponsoring Consortium Open Access Publishing Particle Physics
SOAP	Study of Open Access Publishing (FP7 Project)
STM	Science, Technology, Medicine

Preface

Disclaimer

Parts of this research have been conducted in collaborative large-scale projects. This concerns chapters 2 and 3, where the advantage of the size and diversity of the consortia in the respective projects has been used to obtain large-scale results across disciplines. I have contributed to the development of the research questions, the set-up, realization and analysis. The conception, analysis and interpretation of the respective results presented in this thesis have been conducted by me.

Publications

This thesis is an independent research achievement which has been derived over a period of three years. During this time related articles have been published, all of them with significant, if not leading, contribution by the author of this thesis.

Drivers and barriers to Open Access Publishing

Dallmeier-Tiessen S., Darby, R., Görner, B. et al. (2011). Open Access journals – what publishers offer, what researchers want. Information Services & Use 31 (2011) 85–91. doi: 10.3233/ISU-2011-0624

Dallmeier-Tiessen, S. & Lengenfelder, A. (2011). Open Access in der deutschen Wissenschaft – Ergebnisse des EU-Projekts „Study of Open Access Publishing“ (SOAP), GMS Med Bibl Inf 2011;11(1-2):Doc03. doi: 10.3205/mbi000218

Dallmeier-Tiessen S. Darby, R., Görner, B. et al. (2011). First results of the SOAP project. Open access publishing in 2010. Retrieved from arXiv:1010.0506

Drivers and barriers to research data sharing

Darby R., Lambert S., Matthews B., Wilson M., Gitmans K., Dallmeier-Tiessen S., Mele S., Suhonen J. (2012). Enabling Scientific Data Sharing and Re-use. 2012 IEEE 8th International Conference on E-Science (e-Science). pp.1-8. doi:10.1109/eScience.2012.6404476

Dallmeier-Tiessen, S., Darby, R., Gitmans, K., Lambert, S., Suhonen, J. A., & Wilson, M. (2012). Compilation of results on drivers and barriers and new opportunities. Retrieved from

http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/08/ODE-CompilationResultsDriversBarriersNewOpportunities1.pdf

Case Study in the High Energy Physics Community

Dallmeier-Tiessen, S. & Weiler, H.(2012). Exploring the web as a working space together - a community and its digital library. In Ockenfeld, M., Peters, I., Weller, K., (Eds.). Social Media und Web Science. Das Web als Lebensraum: Proceedings 2. DGI-Konferenz/ 64. Jahrestagung der DGI. (pp. 195-203). Frankfurt am Main, Germany: Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis e.V. ISBN: 9783925474729

Brooks, T. C., Carli, S., Dallmeier-Tiessen, S., Mele, S., & Weiler, H. (2011). Authormagic in INSPIRE Author Disambiguation in Scholarly Communication. In Proceedings of the ACM WebSci'11, June 14-17 2011, Koblenz, Germany. (pp 1-2). Retrieved from http://journal.webscience.org/485/1/158_paper.pdf

Pracyk, P., Nogueras-Iso, J., Dallmeier-Tiessen, S. & Whalley, M. (2012). Integrating Scholarly Publications and Research Data - Preparing for Open Science. A Case Study from High Energy Physics with Special Emphasis on (Meta)data Models. In Dodero, J. M.; Palomo-Duarte, M. & Karampiperis, P. (Eds). Metadata and Semantics Research Conference 2012" (pp. 146-157). doi: 10.1007/978-3-642-35233-1_16

Overarching publications (cited frequently in the introduction chapter)

Dallmeier-Tiessen, S. (2011). Strategien bei der Veröffentlichung von Forschungsdaten. In Büttner, S., Hobohm, H., & Müller, L. (Eds.). Handbuch Forschungsdatenmanagement (pp. 157–168). Bad Honnef, Germany: Bock + Herchen. ISBN: 9783883472836

Dallmeier-Tiessen, S. (2012). Die wissenschaftsorientierte Publikation von Forschungsdaten. In Hohoff, U. & Lülfing, D. (Eds.). Bibliotheken für die Zukunft - Zukunft für die Bibliotheken. 100. Deutscher Bibliothekartag in Berlin 2011 (pp. 75–86). Hildesheim, Germany: Georg Olms Verlag

1. Introduction

1.1 Motivation and aim of thesis

> *"The Internet has fundamentally changed the practical and economic realities of distributing scientific knowledge and cultural heritage. For the first time ever, the Internet now offers the chance to constitute a global and interactive representation of human knowledge, including cultural heritage and the guarantee of worldwide access."*
>
> *Quote from the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*[1]

This citation highlights the strong potential offered by the advent of the internet for an open and global sharing of research materials. Recent studies however have shown that researchers[2] do not yet use the full potential offered by the internet (Carpenter et al., 2012; Meyer et al., 2011; Nentwich, 2009). This is somewhat surprising, as within our private life practices have been changing rapidly in the past decades, with platforms and services enabling open sharing and rapid communication. But within scholarly communication, even though (open) sharing of research materials of any kind is possible, it appears that few researchers actually engage in (open) scholarly sharing activities. In spite of the advent of the internet and the subsequent pervasiveness of new communication channels, the customary workflows in scholarly communication have not changed much from those used in the traditional paper-driven production process. Many new services and communication channels in scholarly communication have emerged (some of them being more successful than others): within some research communities communication channels now comprise blogs, wikis and other Web 2.0 features[3]. But they have not impacted the workflows of scholarly communication across disciplines. It appears that researchers do not use many of the opportunities offered by digital scholarly communication. This applies in particular to Open Access (OA) publishing, research data sharing or Web 2.0 features in scholarly communication (Cullen &

[1] http://oa.mpg.de/files/2010/04/berlin_declaration.pdf [accessed September 12, 2012].

[2] In this thesis the term researchers refers to female and male researchers. However, the male person is used in this thesis to allow for a convenient reading flow. Female and male persons are nevertheless equally addressed.

[3] In this thesis the term Web 2.0 refers to a collection of collaborative and interactive tools that empower users to create content on the Web, and participate in and contribute to services online (cf. Nentwich, 2009). The bidirectional communication flow is important to highlight. Examples in scholarly communication include Open Peer Review processes (cf. Müller, 2009).

Chawner, 2010; Nelson, 2009; Ware, 2008). This shows tools and workflows that facilitate open sharing, communication, discussion and review of scientific results openly are not widely adapted.

This lack of change in the field of scholarly communication, in particular with regard to Open Access publication and research data sharing, is in contrast to an emerging and increasing demand for change and openness by policy making bodies, funding agencies and society[4] as a whole. This demand is reflected in more and more policies that impose openness with regard to publications and data on researchers (e.g. National Science Foundation[5] in the US, National Environmental Research council[6] in the UK, Science and Technology Facilities Council[7] in the UK, Wellcome Trust[8] in the UK).

The reasons for these gaps between current practices and demand need to be understood thoroughly so that strategies can be developed to further the advancement. Even though the framework needs to be considered a dynamic field, it is necessary to understand the underlying drivers and barriers. This thesis provides an in-depth study of those drivers and barriers, and will facilitate the development of better and more tailored support for researchers.

The first part of this thesis investigates the current situation in respect of the drivers and barriers that act on researchers across disciplines when engaging with digital scholarly communication. In the second part of this thesis, these findings are studied in practice and in more detail by means of a case study in the discipline of High Energy Physics (HEP). This case study is also used to further scrutinize the role of information management.

The first research part focuses on a cross-disciplinary analysis of two major innovations in digital scholarly communication: OA publishing and research data sharing. This leads to the following research questions:

[4] The Open Knowledge Foundation, for example, points to "the potential to deliver far-reaching societal benefits […]: Better governance […], better culture […], better research […], better economy […]., see http://okfn.org/about/vision/ [accessed September 10, 2012].

[5] http://www.nsf.gov/bfa/dias/policy/dmp.jsp [accessed September 9, 2012].

[6] http://www.nerc.ac.uk/research/sites/data/policy2011.asp [accessed September 9, 2012].

[7] http://www.stfc.ac.uk/About+STFC/37459.aspx [accessed September 9, 2012].

[8] http://www.wellcome.ac.uk/about-us/policy/policy-and-position-statements/wtx035043.htm [accessed September 9, 2012].

- Within a digital environment researchers can now choose to share or publish their articles by OA means, for example by publishing in OA journals. This is commonly referred to as the "gold" route to OA[9]. The following questions need to be addressed: Do researchers publish by gold OA means? What are the drivers and barriers in gold OA? Why do researchers publish gold OA and why not? What can be derived from the answers to these questions for digital scholarly communication in general?
- Within a digital environment researchers can now share additional materials such as research data more flexibly. Research data are part of the research lifecycle in any discipline. Researchers can submit research data to dedicated repositories, share them via dedicated platforms and link them to an article. These questions need to be asked: Do researchers share their research data? What are the drivers and barriers to sharing data[10]? How can the answers be applied to digital scholarly communication in general?

Both of these communication activities have been described as innovations in digital scholarly communication that have not taken off (e.g. Bell, Foster & Gibbons, 2005; Carlson, 2011; Nelson, 2009). This thesis aims to shed light on the drivers and barriers that have shaped researchers' decisions about how they apply these new technologies.

In the study of OA publishing, researchers' views were sought by means of a survey. In the study of research data sharing, researchers and relevant stakeholders working with them were interviewed. The results are presented in a list of the researchers' drivers and barriers, which is discussed within the wider framework of digital scholarly communication.

By definition, the researcher plays a central role in the research lifecycle, from idea generation to data production to article writing. This leading and central role is also evident in (digital) scholarly communication – in article writing and submission, (peer) reviewing, and publishing. The latter is in particular pronounced in Web 2.0 features and research data sharing: here researchers (today) often act as independent content creators on the web. Thus, the researcher is at the heart of

[9] Today, two main principles of OA are distinguished: the green road and the gold road (see also Suber, 2012). The green road refers to "self-archiving" in a disciplinary or institutional repository. A comprehensive overview of existing repositories can be found in the Open Directory of Open Access Repositories (OpenDOAR). The gold road is defined as primary OA publishing in journals. An overview of OA journals currently available is given in the Directory of Open Access Journals (DOAJ). This thesis focuses on gold OA (in particular in chapter 2); whenever needed the view is extended to the green road. Gold OA and OA publishing are used as equivalent terms in this thesis. A more detailed definition is given in chapter 1.3.2 in this thesis.

[10] In this thesis data sharing and research data sharing are used as equivalent terms.

questions about the research process and the methodologies of scholarly communication that are employed[11].

In the second part of this thesis these results are studied in practice by means of two independent case studies in the HEP community. HEP-specific aspects of the drivers and barriers previously identified are derived. A driver and a barrier are each selected for the case studies. The first one studies the engagement of researchers in using the tools of a large-scale digital library. The results are studied quantitatively. The second case study focuses on research data sharing and uses an embedding situation in the community[12] to study one barrier in more detail. A qualitative approach is taken here.

The case studies allow reviewing the driver and barrier in practices. They are in particular used to investigate the role of information management[13] in the final part of the thesis. It focuses on the support provided to researchers using digital scholarly communication tools and services. The results shall be used to envision the potential future scenarios.

In summary, this thesis aims to improve our understanding of why scholarly communication today has not been as greatly affected by the digital revolution, as has our private life. It will improve our understanding why digital scholarly communication remains focused on the traditional paper-orientated publication process and why engagement with digital scholarly communication services and the practices of Open Science[14] are still lagging behind. Discipline-specific practices are discussed within the framework of drivers and barriers.

[11] It is important to note that the individual researcher is usually embedded in a research community or institution that influences or guides researchers' habits as well. This is reflected in the approach and the respective discussions.

[12] The approach follows an abstracted concept of the embedded librarianship (cf. Kvenild & Calkins, 2011) and embedded research information manager (cf. Walshe, 2011).

[13] See for example Degkwitz & Schirmbacher (2007) for a review of influencing factors for a changing information management at university level in Germany. See Walshe (2011) for recommendations of how research information management could be performed in the future.

[14] Definition of Open Science according to Michael Nielsen: "Open science is the idea that scientific knowledge of all kinds should be openly shared as early as is practical in the discovery process." http://www.openscience.org/blog/?p=454 [accessed August 1st 2012]. Principles for Open Science are given by Science Commons: http://sciencecommons.org/resources/readingroom/principles-for-open-science/ [accessed August 1st 2012]. It needs to be noted that the definition of „open" does vary in the different movements and organisations. Discussions are ongoing to define which materials and research activities are addressed. But, it is evident that OA and research data sharing are part of it.

1.2 Structure of the thesis

Following the general structure (Figure 1), the thesis starts with a description of the framework, the relevant terms and implications for the thesis in chapter 1.3. The focus is on the research lifecycle model, digital scholarly communication, OA, as well as research data sharing. A common terminology is developed, which is needed to facilitate the interdisciplinary scope and understanding of this study.

In the first research chapter drivers and barriers are investigated across disciplines. In the beginning gold OA is studied in detail (chapter 2) as one of the challenges, innovations and opportunities in digital scholarly communication. A quantitative analysis (conducted in 2010) is used to understand drivers and barriers across disciplines in OA publishing today. This comprises a large-scale survey on researchers which highlights major themes which bear on scholarly publishing behavior. While this is a cross-disciplinary study, discipline specific differences are highlighted.

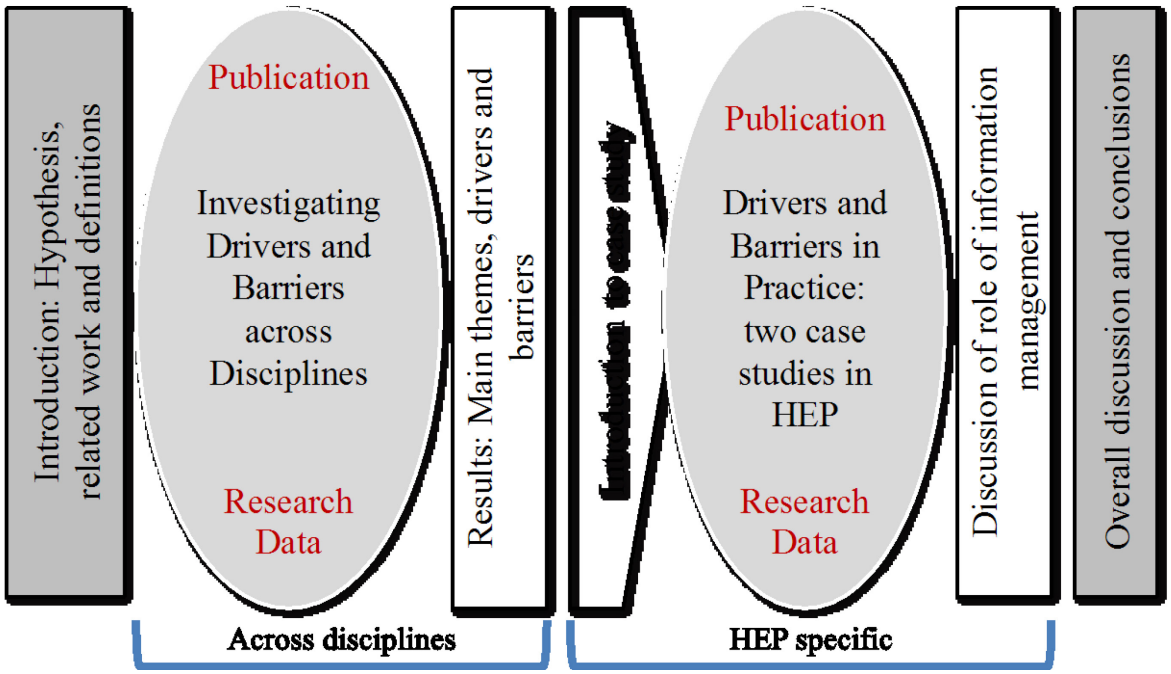


Figure 1: Overall structure of thesis

Next, drivers and barriers in research data sharing are investigated across disciplines (chapter 3). Digital scholarly communication facilitates a new handling of the research data digital object, its preservation and sharing. Drivers and barriers in research data sharing today are investigated using a two-step qualitative approach.

Both research chapters explicitly study across disciplines with an eye on disciplinary practices. Finally, the results are discussed in the wider framework of digital scholarly communication and the participation of researchers therein.

In the second part of this thesis the results of the analysis of drivers and barriers in digital scholarly communication from the first chapters are taken into the community of HEP (chapter 5). A brief introduction in the HEP community and its specific framework in digital scholarly communication are given. In addition, specific aspects concerning drivers and barriers are highlighted. Based on the results of the first research part, two themes have been selected for these instrumental case studies in the HEP community. This approach allows more practical insights into the themes and will provide insights on the potential role of information management in digital scholarly communication. The themes to be studied are "reputation" and "hesitation" as part of the societal layer.

The first case study (chapter 5.4) focusing on publications is implemented in the large scale digital library INSPIRE[15]. The reputation driver is studied via a dedicated engagement strategy with the large HEP community. The second case study focusing on research data is given in chapter 5.5. The description of a close collaboration of the author of this thesis with the HEP community underlines the strategic development of data preservation and access to data in this discipline. It is shown how the hesitation barrier can be overcome. In practice, this strategy results in tailored services for data sharing through the digital library service INSPIRE.

The thesis concludes with a discussion of the drivers and barriers in digital scholarly communication (chapter 5.6). Finally, the results are discussed with regard to their applicability across disciplines (chapter 5.7). Furthermore, they are discussed in relation to the role of information management and science (i.e. in community engagement), in chapter 5.8.

1.3 Background to research questions

This chapter provides an overview of the framework of the thesis. The concepts used and built upon in this thesis are defined, explained and set into the context of the research questions. First

[15] INSPIRE is the main digital library serving the HEP community. It is run by the four main laboratories in HEP: the European Organization for Nuclear Research (CERN), Deutsches Elektronen-Synchotron (DESY), Stanford Linear Accelerator Laboratory (SLAC) and Fermilab. It is accessible via http://inspirehep.net/ [accessed July 12, 2012]. More details are given in chapter 5.4.

scholarly communication, its tradition, role and stakeholders as well as its transition to the digital environment are described. Briefly related works on the subject of drivers and barriers in digital scholarly communication are discussed.

In a second step this introduction focuses on the research questions and elaborates the concepts of OA and research data sharing. These are the two major phenomena investigated in the framework of digital scholarly communication. Both are central to the first research part of this thesis. By defining these two concepts and by developing a common terminology a consistent research approach in the following is facilitated. Existing related works on drivers and barriers are discussed in the introduction of the individual research chapters 2 and 3.

1.3.1 Scholarly communication and digital scholarly communication

The tradition of scholarly communication has been in existence for centuries. The Philosophical Transactions (of the Royal Society) was the first academic journal, in its modern sense founded in 1665 (Spier, 2002). Since then, scholarly communication has been focused on the idea of exchanging scientific expertise and preserving it. Ware & Mabe (2009) name four functions of a journal:

- Registration: ownership of an idea
- Dissemination: communication of findings
- Certification: ensuring quality control
- Archival record: preservation of a record for future reference

Roosendaal & Geurts (1997) point to a fifth function: the reward system. This is based on the citation of previous works and corresponding metrics.

Disciplinary differences have existed throughout the history of scholarly communication, i.e. with dominance of book publishing in some fields in the Humanities and Social Sciences (Williams et al., 2009).

Publishing research in journals and proceedings has undergone various phases and transitions through time, mainly linked to the diversification of research fields, the development of individual community practices, the development of a competitive scholarly communication business and the advent of the internet and associated online publishing opportunities (Van De Sompel et al., 2004; Ware & Mabe, 2009). In that regard it is important to note the growing body of publications (cf.

Weiler, 2012) over time, and in recent years the increasing importance of publications for purposes of research evaluation and funding allocation.

Researchers, libraries, publishers and service providers have established roles in this domain, sharing responsibilities for particular functions, such as quality assurance, distribution and preservation of knowledge. Roles and responsibilities have also been changing, i.e. as distribution channels are not solely focused on print publications distributed via libraries anymore, but are open to any stakeholder who wants to share materials online.

The term "digital scholarly communication" refers to scholarly communication in such a digital environment. Maron & Smith (2009) identify eight different types of digital scholarly resources: e-only journals; reviews; preprints and working papers; encyclopedias, dictionaries, and annotated content; data; blogs; discussion forums; professional and scholarly hubs.

With the advent of the web, new tools and services could change workflows and dissemination channels within digital scholarly communication. And indeed, changes are visible: within the STM (Science, Technology, Medicine) domains articles are now submitted to journals online, peer review is usually conducted via dedicated web services, and publishers usually offer a digital version (or e-only version) of articles. Furthermore, the digital environment allows for an enhanced incorporation of additional materials, such as slides, data, and code into the scholarly article (Dallmeier-Tiessen, 2012). In spite of these changes, one of the basic principles of scholarly communication has remained the same since the beginning: peer review. This concept implies the quality assurance procedure through peers in the same research domain. Today, different occurrences of peer review exist (cf. Müller, 2010) also incorporating additional Web 2.0 features (such as post-publication commenting in some Public Library of Science (PLOS) journals[16]), but the basic principle remains the same.

However, some studies have investigated the slow uptake of Web 2.0 features by researchers. Collins & Hide (2010) examine the usage of Web 2.0 tools by researchers. They conclude that "[…] it is reasonable to suggest that overall use of information sharing Web 2.0 tools is by no means intensive among researchers". Furthermore, they highlight that "[o]verall it appears that researchers are not engaging systematically with Web 2.0 tools."

[16] http://www.plos.org/ [accessed August 1, 2012]

Similar results have also been presented by Procter et al. (2010). Their study shows that "[a]doption of Web 2.0-based novel forms of scholarly communication has reached only modest levels so far". They describe the usage as being "still in a rather fragmentary manner". But the authors of this study also showcase the existence of more successful models, which "appears to revolve around more dispersed and dynamic innovation patterns arising from community-based activities and from start-ups".

Hurd (2000) does see a change in roles with the advent of digital scholarly communication. "Scientists become publishers" so that roles are "blurred". The author points to the possibilities of publishing via websites or other services on the Web. First studies are available on the usage of such tools and workflows (e.g. Björk et al., 2010). However, it is not clear how intensively researchers use these opportunities (Hurd, 2000) and what are the drivers and barriers. The change of roles and associated responsibilities is important to stress. By definition a Web 2.0 environment, in scholarly communication as in other domains, moves beyond information "viewing". It focuses on a participatory Web that facilitates "sharing" – also in respect of new digital objects such as research datasets. The handling of such materials or the participation in an innovative (possibly open) peer review process requires that researchers take an active role beyond paper publication.

This new role of researchers becomes evident when looking at the research and knowledge production processes (Figure 2). Today, it is possible to integrate all processes in a digital environment, from idea generation to collaborative processes, data generation, data sharing and publication. Researchers do play a central role in providing such content online in the respective frameworks, but even more, they can also enrich the content by providing more content information. In practice this means, for example, linking related materials, publications, slides, documentation or research data.

But profound changes in digital scholarly communication practices and models remain restricted to particular domains and specific services (Maron & Smith, 2009). It appears that workflows and practices have been mirrored from the paper publishing environment to the digital environment without profound changes, for example, with regard to peer review. As Mulligan & Mabe (2011) state: "[…] the fundamentals of formal scholarly communication steadfastly remain the same"; and "[…] none of the advances […] have yet to affect the fundamental form of the formal scholarly article".

In the same study, Mulligan & Mabe postulate: “it is clear that, [however] the scholarly communication system develops in the coming years, it will only be successful if it satisfies the needs of the researchers and minimizes the time spent on preparing materials for publication as well as the time spent finding and retrieving articles”. This is particularly important in respect of this thesis, which aims at an improved understanding of current drivers and barriers in this highly dynamic environment. In achieving this, it should provide a solid knowledge base for information management science to support researchers in digital scholarly communication.

Fry & Talja (2007) state that the audiences and alliances for web production, publishing, and related activities have not radically changed from the print world. Interestingly, they point out that disciplinary differences occur. They highlight “that communication systems designed for one discipline can prove inappropriate for, or even harmful to, another”. The authors point out that “the focus on disciplinary cultures as the factor explaining differences helps us to reach a closer understanding of what lies at the heart of the shaping of the networked environment”. Thus, there is an urgent imperative to understand the needs of the research communities, i.e. to understand the relevant drivers and barriers existing in the current system. With a better understanding, information management will be able to better support the transition of the communities to digital scholarly communication. Thus, the second research part in this thesis focuses on a disciplinary case study to review drivers and barriers in digital scholarly communication in practice.

In summary, the new opportunities offered by digital scholarly communication are widespread, but two innovations have gained widespread international attention: OA to publications and OA to data. Both are investigated in depth in this thesis. In the following the necessary frameworks for both phenomena are given.

1.3.2 Open Access

Open Access in scholarly communication generally describes free access without restrictions to scholarly materials; currently the discussion mainly concerns OA to publications and research data, but it is part of a more general “Open Science” movement. Open Access has been facilitated through the new opportunities in digital scholarly communication, using new (digital) preservation, distribution and dissemination channels.

The Budapest Open Access Initiative in 2002[17], the Bethesda Statement on Open Access Publishing in 2003[18] and the Berlin Declaration on Open Access to Knowledge in the Science and Humanities[19] are the landmark declarations that have been attracting international attention and signatories in support of OA since the movement was launched.

The Budapest Open Access Initiative defines OA as the freedom to "read, download, copy, distribute, print, search or link to the full texts of articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal or technical barriers other than those inseparable from gaining access to the Internet itself". Peter Suber, a professor at Earlham College, defines OA literature as "digital, online, free of charge, and free of most copyright and licensing restrictions"[20].

OA literature has a long tradition in the discipline of HEP. In this domain, a preprint culture alongside journal publishing was established as early as the 1950s (Goldschmidt-Clermont, 1965; Aymar, 2009). Due to the long processing times in journal publishing this parallel distribution channel developed. One of the main tools currently being used for preprint distribution in HEP is the arXiv[21] repository, which has been used since 1991 to upload and distribute preprint versions of articles (Ginsparg, 2011). This development is considered the beginning of (the green road to) OA. But the technical precondition for OA in a general sense was the advent of the internet, which facilitates direct online access to research results from a researcher's own workspace.

Today, two main principles of OA are distinguished: the green road and the gold road. The green road refers to "self-archiving" via disciplinary repositories (such as arXiv) or institutional repositories (such as CERN Document Server[22] or edoc-server[23] of the Humboldt Universität zu Berlin). Numerous repositories are available to support the green road to OA. The fact that different versions of papers known as preprints and postprints are archived indicates that research materials may be shared at different states in the research lifecycle: researchers might share a copy of their article in a repository once it is drafted or once it has been published in a journal – or at any

[17] http://www.soros.org/openaccess/read [accessed August 1, 2012].

[18] http://www.earlham.edu/~peters/fos/bethesda.htm [accessed August 1, 2012].

[19] Ibid. 1 [accessed September 12, 2012].

[20] http://www.earlham.edu/~peters/fos/overview.htm [accessed August 1, 2012].

[21] http://www.arxiv.org [accessed July 30, 2012].

[22] http://www.cds.cern.ch [accessed July 30, 2012].

[23] https://edoc.hu-berlin.de/ [accessed July 30, 2012].

time in between (depending on the terms publishers attach to submission and publication in their journals). This can involve additional challenges in respect of licensing. Registries for repositories have been developed (e.g. Open Directory Open Access Repositories[24] or Registry of Open Access Repositories[25]) to provide guidance for the individual stakeholders.

This thesis focuses mainly on gold OA which is defined as primary publication in OA journals. An overview of OA journals currently available is given in the Directory of Open Access Journals (DOAJ[26]), which currently lists 7811 journals (as of May 30th 2012). In opposition to the traditional subscription-based journals, OA journals provide immediate free access without charge to scholarly content. Business models can be mixed, but include revenue streams from article processing charges, membership fees, advertisements, sponsorships, subscription to or sale of print versions. In addition, many publishers provide a hybrid model (see also Suber, 2012), which enables researchers to pay for OA publication of a particular article in subscription journals. Individual funding and support mechanisms have been developed by some funding bodies and institutions. More details on gold OA are provided in the introduction of chapter 2.

Alongside the definition of OA, themes such as licensing are discussed. In opposition to previous publishing practices, publishing by OA standards implies that copyright stays with the author and reuse is facilitated via dedicated licenses[27]. Since the declarations mentioned above, OA has taken off as a cross-disciplinary and overarching topic in digital scholarly communication. According to Müller (2009) there are three reasons for the greater awareness and motivation in respect of OA: overcoming the crisis in journal publishing, progress and freedom in research, and global fairness. This thesis will shed further light on the motivations and barriers in regard to (gold) OA.

Results from Rowlands & Nicholas (2005) show that "researchers are rapidly becoming more informed about OA publishing and institutional repositories." The authors conducted a survey of journal author behavior and attitudes in 2005, which highlighted a significant positive shift in awareness in comparison to their previous survey conducted in 2004.

[24] http://www.opendoar.org/ [accessed June 1, 2012].

[25] http://roar.eprints.org/ [accessed June 1, 2012].

[26] http://www.DOAJ.org [accessed May 30, 2012].

[27] See the Budapest Open Access Initiative for details. It is to be noted that exceptions exist, e.g. particular publishing houses use other licences.

In practice researchers now have the opportunity to enhance their publishing process by OA means either via the green or the gold route. This means they can choose to share their articles at any time through a repository while submitting to a traditional journal, and they could submit their articles to an OA journal directly. Chapter 2 will outline that this is not (yet) common practice and the research chapter investigates the details of drivers and barriers accordingly. This is particularly needed as the awareness and demand in the society and on the policy-making side is increasing. The resulting demand for OA to scholarly materials is especially evident in the case of (national) funding bodies, e.g. the European Commission, which demands OA publication as standard for the results of research it has funded (European Commission, 2012a). Increasingly, funders link funding provisions with requirements to publish the results by OA means. An overview of such policies can be found in the Registry of Open Access Repositories (ROAR[28]). The recent statements by the European Commission (2012a) also point out that there is a need to move beyond OA to publications. There is a pressing demand for public access to research materials such as research data.

1.3.3 Research data and research data sharing

"Research data" refers to a primary product of the research lifecycle. A definition is given by the Committee for a Study on Promoting Access to Scientific and Technical Data for the Public Interest (1999), *„Data* are facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors. A data element is the smallest unit of information to which reference is made."As part of the knowledge production process data are produced in all disciplines (Figure 2). Data are not a new object in research. Research data have always been part of the process, but can now be handled differently as digital objects. Traditionally research data was preserved in non-digital formats, such as charts, tables or graphs, or tapes. Research data as a digital object or representation can now be stored and integrated differently in the research lifecycle and in digital scholarly communication (cf. "Principles for Open Sciences"[29] and Dallmeier-Tiessen, 2012).

[28] http://roarmap.eprints.org/ [accessed July 13, 2012].

[29] http://sciencecommons.org/resources/readingroom/principles-for-open-science/ [accessed July 13, 2012].

Research data might be produced and used throughout the research lifecycle[30] depending on the discipline. The role of data in the research process differs according to the research domain, and data handling varies, with different steps that might apply such as data collection, processing, re-processing etc. Accordingly also research data definitions vary. Even though there is a common agreement on the usage of the overall term "research data", some also refer more specifically to raw data, primary data, secondary data, etc. The latter usually refers to more advanced processing steps in the individual disciplines. In general it can be said that research data is analyzed to answer specific research questions and to browse for new research ideas. Data can be shared at any time during the research data production process – immediately after production or after an embargo or study period (Birney et al., 2009; Dallmeier-Tiessen, 2011; Schofield et al., 2009).

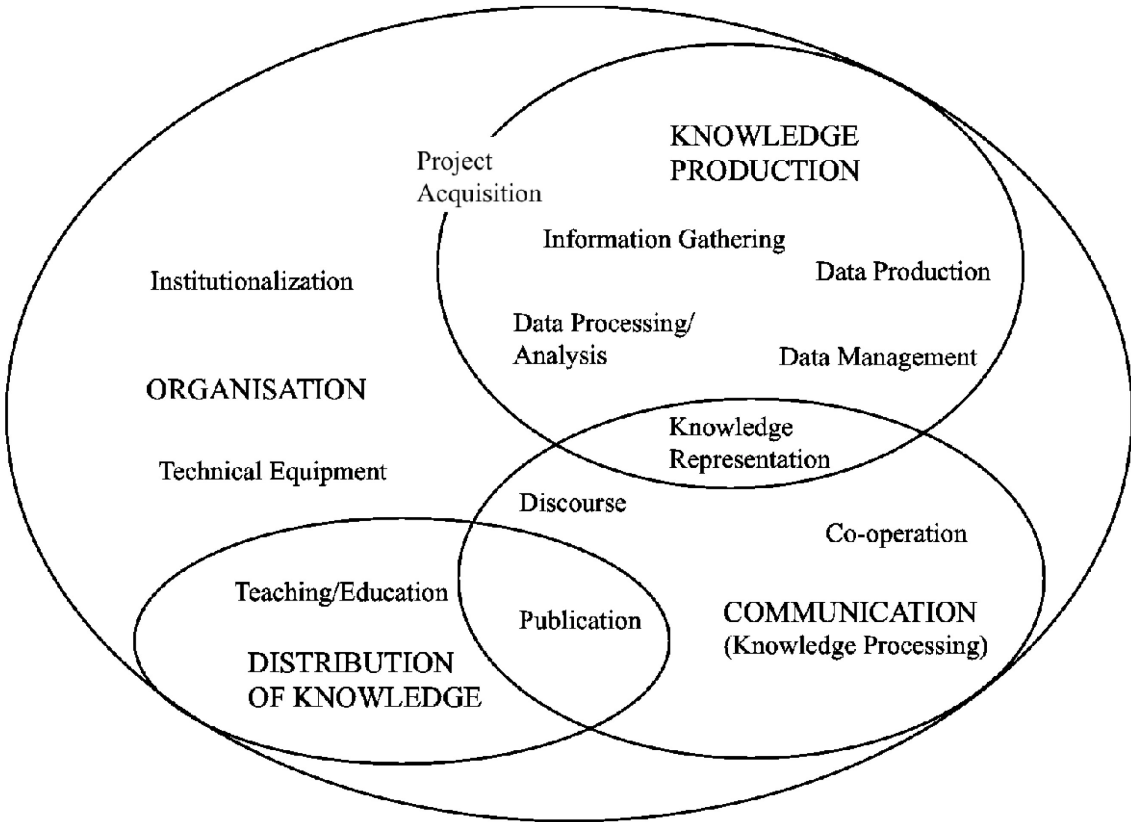


Figure 2: Research data is an integral part of the knowledge production process (Figure modified after Nentwich, 2003)

Dealing with digital research data in scholarly communication is a relatively new challenge that has been addressed by different stakeholders. Several relevant aspects in particular have been

[30] Research lifecycle as described for example by the Joint Information Systems Committee (JISC), http://www.jisc.ac.uk/whatwedo/campaigns/res3/jischelp.aspx [accessed September 5, 2012].

studied, such as data preservation, management, access models, quality assurance procedures and workflows for data sharing (Büttner, Hobohm & Müller, 2011; Pryor, 2012). In this thesis data sharing is studied. By its very nature this touches parts of the other aspects, but emphasis is given to researchers' participation in data sharing.

The term "data sharing" describes the process of making research data available for multidisciplinary purposes, whether for future reuse or reinterpretation (Borgman, 2010, 2012; "Data's shameful neglect", 2009; Nelson, 2009; Tenopir et al., 2011). Such data can be of interest for a wider laymen audience, other research communities or colleagues for example. Data sharing commonly refers to making research data publicly available (by OA means). It can take place on a global or more restricted level, for example on community platforms or with access restrictions (for a number of possible reasons). In this thesis, the term data sharing refers to making data available for future purposes, such as reuse and reinterpretation, and thus also comprises research data that is not published by OA means, but is shared within the community, with colleagues or within the research project (cf. Borgman, 2010). Nevertheless it is important to highlight that researchers or groups of researchers, as well as communities, make individual decisions about where, when and how to share research data (see also Dallmeier-Tiessen, 2011).

In addition, the term "open data" is often used to describe the unrestricted access to data and entails expectations as to licensing data (see the Panton Principles for Open Data in Science[31], Protocol for Implementing Open Access Data[32]). The term does not only focus on research data, but is also extended to government data, for example. "Linked open data" describes a standardized structure to open data (cf. Heath & Bizer, 2011). This approach to data description in particular facilitates the discoverability and reuse of research data; in particular in regard to large-scale data sharing.

Discipline-specific characteristics also apply to the individual standards, size and complexity of research data sets. Habits and practices with regard to research data sharing appear to be very discipline-specific as well (Tenopir et al., 2011). Different models exist and emerge, including dedicated data repositories and dedicated data journals. Previously published research data have often been attached to a publication, mainly as supplementary materials, and as tables, graphs, and plots. Nowadays, with the digital environment data publication options are more flexible (cf. Dallmeier-Tiessen, 2011). But it is not only the models that differ within a discipline: the nature of progress is also different. Within the molecular biology early community agreements (e.g. the

[31] http://pantonprinciples.org/ [accessed July 26, 2012].

[32] http://sciencecommons.org/projects/publishing/open-access-data-protocol/ [accessed July 22, 2012].

Bermuda Principles from the Human Genome Project[33]; Smith & Carrano, 1996) contributed to a wide adoption of data sharing as a common practice (Wellcome Trust, 2003).

Such advances in data sharing have raised international awareness, also beyond disciplinary borders (cf. Science, 2011). There is an increasing public demand for OA to research data by policy makers as well as funding bodies. In a survey launched by the European Commission (2012b), 90% of the respondents agreed/agreed strongly that "research data […] that results from public funding, should, as a matter of principle, be available for reuse and free of charge on the internet". Recent results (outlined in chapter 1.1) show that this is currently not the case.

According to Borgman (2012) there are four main reasons to share data:

- To make the results of publicly funded data available to the public,
- To enable others to ask new questions of extant data,
- To advance the state of science,
- To reproduce and verify research.

Funders and society increasingly demand long-term data preservation and permanent access to research data. Strategic data management and OA to data will improve further use and reuse of the datasets being produced, and thus might also contribute to a more economic usage of public money (Kroes, 2010). Moreover, this is seen as a matter of transparency - allowing for long-term preservation of the integrity of the research projects and results. Thus, funders' requirements usually focus on several aspects in respect of research data: data management plans[34] (which are increasingly demanded for proposal submission), data preservation, and data access models. For example, the National Science Foundation (NSF)) requires data management plans upon proposal submission[35]. Among other things they require specification of preservation and data sharing plans. Similar policies or recommendations are issued by many funding bodies, with the number increasing steadily.

In Germany too, several initiatives in the same vein have emerged during recent years. For decades the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation[36]) required the

[33] http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml [accessed July 22, 2012].

[34] For example http://www.nsf.gov/eng/general/dmp.jsp [accessed September 14, 2012].

[35] http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp [accessed July 24, 2012].

[36] http://www.dfg.de/en/dfg_profile/mission/index.html [accessed September 13, 2012].

preservation of data for 10 years after production (Deutsche Forschungsgemeinschaft, 1998), but did not make any specification in respect of access, nor any data management plans. In 2009 more specific recommendations were published, highlighting metadata and quality assurance measures, for example (Deutsche Forschungsgemeinschaft, 2009). In 2010 the DFG published recommendations for researchers applying for grants to describe measures to preserve data and facilitate data reuse (Deutsche Forschungsgemeinschaft, 2010; Pampel & Bertelmann, 2011). Recently overarching initiatives have emerged, such as the "Priority Initiative Digital Information of the German Research Organizations", which published "Principles for the Handling of Research Data" (Allianz der deutschen Wissenschaftsorganisationen, 2010). The recent statements in 2009 to 2011 reflect an increased significance of permanent OA to research data over time.

In such policies and position papers the demand for enhanced data management indicates a growing awareness of data preservation and data sharing as separate but linked topics. This thesis focuses on the latter. Several stakeholders participate in discussions about data sharing: libraries, publishers, research organizations, infrastructure providers[37], representatives of research communities, and active researchers.

1.4 Definitions

Based on this introductory chapter, a brief definition of the main terms is given here. These definitions will facilitate a common understanding in the following chapters[38].

- Digital scholarly communication: the process of scholarly communication in a digital environment, highlighting the opportunity to integrate new innovative services, tools and digital objects into the communication of scientific results.
- Gold OA: immediate, unrestricted and global access to articles published in OA journals[39]. Open Access models include an appropriate licensing of the content that allows further (re)use of materials.

[37] Such as local and global data centers and repositories (depending on the discipline). This term comprises e-infrastructures as well.

[38] It is not the scope of this thesis to elaborate and discuss the definitions of these terms in detail, but rather to analyze attitudes and practices with regard to the individual topics. For this reason the definitions given in this chapter are concise. More on the individual topics and related research are provided in the individual chapters and discussions.

- Research data sharing: making research data available for future purposes (e.g. reuse and reinterpretation), e.g. via data repositories, as supplements to articles. For the purpose of this thesis controlled and restricted access are both included (as the data sharing takes place within a defined community).[40]
- Publishing: refers to the process of making a scholarly object available within scholarly communication. This is often a shorthand for the publishing of mainly text-based articles in academic journals. Within digital scholarly communication and for the purpose of this thesis the term is used for the publishing of any kind of scholarly material on the Web as well, including research data, blogs, and annotations. This means that data publishing and data sharing are used synonymously in this thesis. As mentioned before, to focus the discussion, data sharing within a closed community (with access restrictions) is also included.
- Publication: refers to the item that is being shared/published, either openly or not. Traditionally this has been a text publication; within the digital environment this term could also apply to other scholarly materials, such as research data, slides, documentations, etc., which could be published online. Bearing in mind exceptions due to disciplinary differences and the existence of emerging scholarly objects, such a product is often quality-assured, either by peer review or other means. Within this thesis, a publication refers to a text publication ("paper"). Whenever applicable the term specifies the object being discussed, e.g. data publication when a dataset is published.

[39] In addition, the green road to OA focuses on "self-archiving", the preservation and distribution of articles via repositories. This approach is considered in parts of the discussion in this thesis and highlighted.

[40] It has to be noted, that research data, data and data sets are used synonymously in this thesis. In this thesis they refer to the primary product of research as explained above (chapter 1.3.3). Data sets contain a connotation on their complexity (cf. Borgman 2011).

2 Drivers and Barriers to Open Access Publishing

2.1 Introduction

Within the framework of digital scholarly communication it is now possible to use new digital publishing models. One of the innovations is gold OA[41], which is studied in detail in this chapter.

Some studies have already investigated the share of OA articles relative to the overall number of scholarly articles available. In 2010 Björk et al. reported that in 2009 8.5% of the peer reviewed articles were freely available at the publisher' site[42], while for an additional 11.9% a free manuscript version could be found. The authors highlighted differences across disciplines and singled out a 6.6% gold OA share of all ISI journals. Laakso et al. (2011) studied the development of OA from 1993 to 2009. They defined three stages of gold OA adaption, which resulted in a 7.7% share of articles being published in OA journals in relation to all peer-reviewed articles in 2009.

The studies above mention disciplinary differences, but so far the reasons why some research communities seem to be rather reluctant to use OA have been mainly based on anecdotal evidence. In respect of gold OA only very little (up-to-date) data has been available that could provide evidence about researchers' attitudes towards this new opportunity in digital scholarly communication[43]. This is particularly true when searching for data that facilitates a comparative review across disciplines (in 2010).

In 2004 Swan & Brown conducted a survey to gather evidence about this issue (with 311 responses) as did Rowland & Nicholas in 2005 (with 5,513 responses). Taking the dynamics of digital scholarly communication into account both are now out of date; but the latter survey did highlight that senior researchers are rapidly becoming more informed about OA and institutional repositories. The authors of the 2005 study highlighted disciplinary practices and national frameworks as important factors affecting author attitudes.

In 2009 Morris & Thorn conducted a survey across disciplines (1,368 researchers, predominantly from the biological sciences, and UK-based). The authors reported that many respondents were in

[41] See page 10 for details.

[42] Björk et al. use a random sample of 1837 titles in their study and a web search engine.

[43] With regard to the green OA model, repository managers describe that researchers often do not share their pre- or postprints online (e.g. Davis & Connolly, 2007; Nicholas et al., 2012; Pelizzari, 2004).

favor of gold OA, but many had concerns as well. Dominant concerns were about "the cost to authors, possible reduction in quality, and negative impact on existing journals, publishers and societies".

Creaser et al. (2010) analyzed a survey among more than 3,000 researchers from Europe in 2009 (supplemented by evidence from focus groups). They highlighted disciplinary differences in awareness of both, the gold and green OA models. In order to identify motivating factors Warlick & Vaughan (2007) conducted 14 semi-structured interviews with researchers. They conclude that for researchers "publication quality" is important for choosing publication outlets. In the selection of OA journals they find that free access and visibility are the main drivers.

This brief summary shows that in 2010 there was only limited knowledge about what researchers think and do about OA, with no comprehensive overview of the current situation: do researchers publish OA or not? And what are the drivers and barriers to OA publishing? An up to date knowledge base is needed to develop a detailed understanding. This will give an interesting perspective on drivers and barriers in publishing research openly in digital scholarly communication.

The research described here was conducted as part of the "Study of Open Access Publishing" (SOAP[44]) project with 6 European partners, including libraries, publishers and research organizations: CERN, Max Planck Digital Library of the Max Planck Society[45] (Germany), Science Technology Facility Council[46] (UK), BioMed Central[47] (UK), SAGE Publishing[48] (UK) and Springer Science and Business Media[49] (Germany). Based on the need for an up to date evidence base on OA publishing and the opportunity to conduct a large scale survey on gold OA via the publishers' and research organizations' networks, the project decided to follow a quantitative approach to study the drivers and barriers in detail across disciplines. The research was conducted and analyzed as a joint effort. The author of this thesis contributed significantly to the survey design, the continuous analysis of the ongoing survey and the final research and

[44] The project was funded for 2 years by the European Commission' s 7th Framework Programme, more details via http://www.project-soap.eu [accessed July 22, 2012].

[45] http://www.mpg.de/en [accessed July 22, 2012].

[46] http://www.stfc.ac.uk/ [accessed July 22, 2012].

[47] http://www.biomedcentral.com/ [accessed July 22, 2012].

[48] http://online.sagepub.com/ [accessed July 22, 2012].

[49] http://www.springer.com [accessed July 22, 2012].

interpretation during the SOAP project. The analysis presented in this chapter has been designed and composed for the purpose of this thesis. The conception and interpretation of the analysis presented here is an independent achievement by the author of this thesis (based on the published SOAP dataset, Dallmeier-Tiessen et al., 2011a).

2.2 Approach

In order to investigate the drivers and barriers to OA publishing (gold OA), the project consortium decided to conduct a large-scale survey. Within the SOAP project the survey was designed and possible distribution channels were identified. It was decided to conduct an online survey with a sampling frame, covering all disciplines and thus allowing for a comparative analysis (cf. Pickard, 2007).

The response pattern was checked and reviewed daily in order to achieve a representative share of the whole research community with respect to disciplines, seniority and origin (country). Within the survey different roles were distinguished: researchers, librarians and publishers. For this thesis, only the researchers' answers are taken into account. The survey consisted of several consecutive parts:

- Demographic information about the researcher, such as age group, research field, type of institution (e.g. university, hospital, research centre…), and country,
- Researchers' experience in scholarly communication in general,
- Researchers' beliefs with regard to OA,
- Researchers' experience with regard to OA,
- Researchers' agreement and disagreement with OA myths.

The majority of the 23 questions were closed (multiple choice); six questions offered in addition a free text box in which researchers could discuss their opinion or experience. In addition, 2 questions asked for a rating of statements (presented to the participant in randomized order)[50].

The survey was set up with the web service Survey Monkey[51] and email invitations were designed. In order to reach all disciplines the survey was distributed via the distribution channels of the

[50] A full overview of the list of questions and the dataset is published in Dallmeier-Tiessen et al. (2011a and 2011b).

[51] http://www.surveymonkey.com/ [accessed May 18, 2012].

partnering publishers in the SOAP project (estimated number of recipients: SAGE 800,000; Springer 250,000; BioMed Central 170,000), plus additional recipients via the members of the Open Access Scholarly Publishing Association (OASPA) and mailing lists of the participating research organizations[52]. In addition, some communities that were known to be underrepresented in the main communication channels were targeted directly via dedicated mailing lists. The individual distribution channels were captured via different collectors in Survey Monkey.

Data quality and progress checks were done daily using the basic tools of Survey Monkey. In disciplines in which a low response was observed, an additional mailing was launched using a sample of addresses obtained from Thomson Reuters three months after the initial launch of the survey (70,000 recipients). In all the survey was live from April 28th to November 15th 2010 (Figure 3).

The final data processing and the analysis of the survey were done with the IBM SPSS[53] software package and Microsoft Excel. In addition, for the analysis of the free text answers, the software package Provalis Research QDAMiner[54] was used for annotation. Within the SOAP project each free text answer was tagged so that a quantitative analysis of significant terms was possible.

The dataset has been made anonymous and published alongside the first publication (Dallmeier-Tiessen et al., 2011a and 2011b). For the anonymization the dataset has undergone several processing steps which are described in the data manual (Dallmeier-Tiessen et al., 2011b). This dataset is used for the analysis presented in the following chapter.

For a thorough validation of the survey, error estimates and potential bias have been studied within the SOAP project. The participation in the survey was high, leading to more than 35,000 responses by researchers. The majority of the responses to the survey was triggered through the distribution channels of the publishing houses Sage, Springer and BioMed Central. The number of responses received allows for an analysis across disciplines. The response pattern has been reviewed with

[52] This included also a multidisciplinary mailing list for the project coordinators and Marie Curie alumni by the 7th Framework Programme.

[53] http://www-01.ibm.com/software/analytics/spss/ [accessed May 18, 2012].

[54] http://www.provalisresearch.com/QDAMiner/Qualitative-Software.html [accessed January 19, 2012].

regard to the seniority, discipline and country distribution[55]. Accordingly, the above-mentioned Thompson Reuters mailing was compiled to compensate for potential underrepresentation.

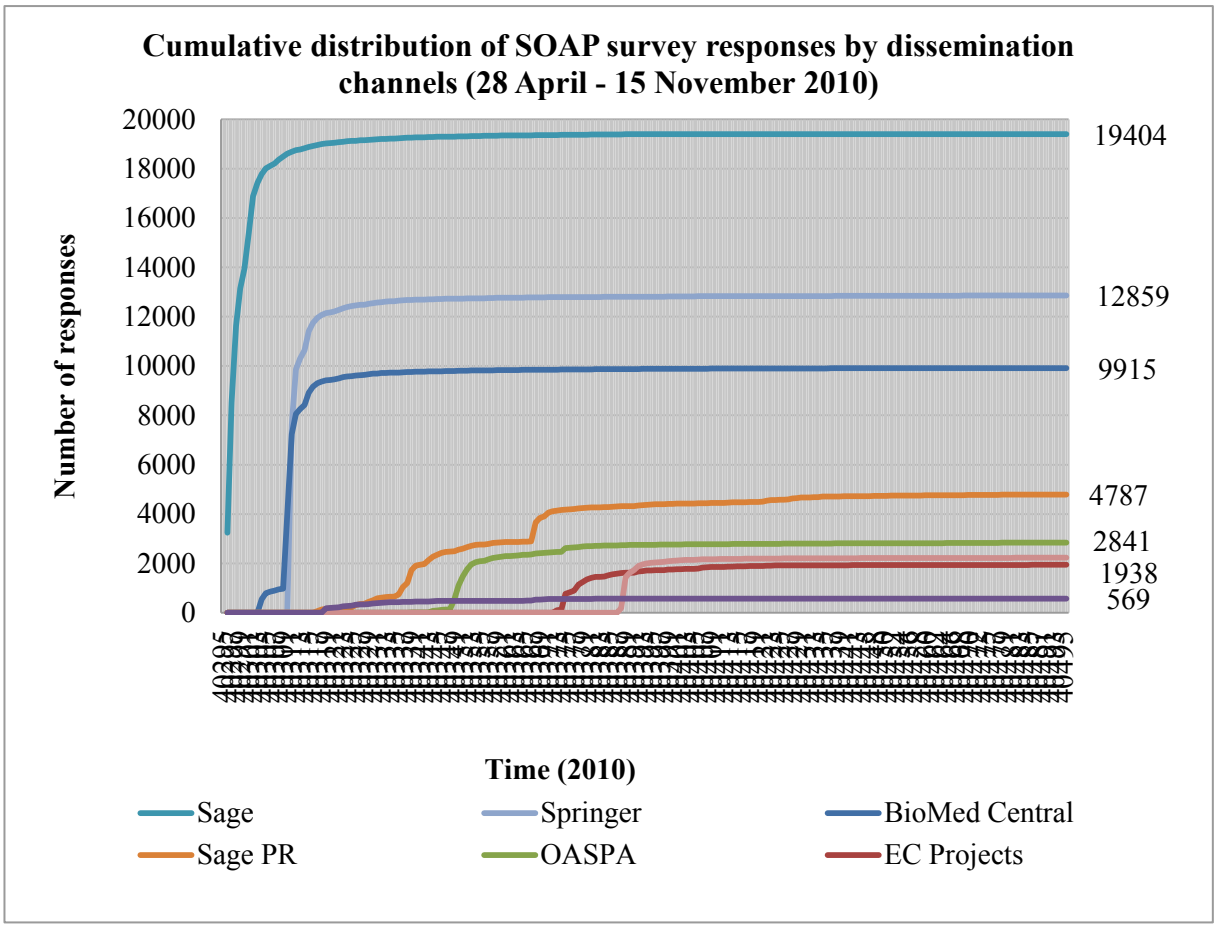


Figure 3: Survey response over time. The plot distinguishes the different dissemination channels. The majority of the respondents were reached via the publishers' dissemination channels (BioMed Central, Sage, and Springer). The survey was also distributed via the OASPA mailing list and other cross-disciplinary resources. It is expected that recipients passed the invitation on to other interested stakeholders. Specific disciplines were targeted with the Thomson Reuters mailing after 3 months of survey distribution. This cumulative plot includes all respondents (researchers, librarians, publishers).

[55] The influence of a possible bias due to "experienced" OA publishers addressed via BioMed Central has been tested. Extracting their responses does not yield statistically significant different results.

In the first and main survey participants were asked if they were willing to participate in follow up questions. Thus, a second survey was launched as part of the project to follow up some of the issues raised in the main survey.

For this thesis, a selection of questions has been chosen in order to provide a comprehensive overview of researcher's drivers and barriers to gold OA. The numbers shown in the figures are given in Appendix A.

2.3 Results

For this brief analysis of drivers and barriers only answers from researchers have been taken into account. In addition, only researchers with experience in scholarly communication have been chosen, selected by having published at least one peer reviewed article (question 12 in the survey). This results in a total of 37,100 responses analyzed in this thesis. In respect of disciplines, there is a dominance of the biological and medical sciences (Figure 4). Almost all countries are represented, but there is a dominance of OECD countries[56].

This chapter will first describe the existing situation by selecting and highlighting questions that describe the state of the researchers' perceptions and actions in respect of OA publishing in 2010.

To begin with, current researchers' mindset and experience concerning OA publishing needs to be studied. Question 9 of the survey ("Do you think your research field benefits, or would benefit from journals that publish Open Access articles[57]?"; see Figure 5) reveals that the vast majority of the researchers finds OA beneficial for their research field (on average 89%). Even though this is the case for all the disciplines studied, strong differences between them are observed. Strong differences also occur on the regional layer, between countries the researchers work in[58]. The free

[56] See also, Simon Lambert for the SOAP consortium, http://www.slideshare.net/ProjectSoap/soap-symposiumtalkii [accessed September 12, 2012]

[57] The definition of OA (gold) as given in the survey: "Many of the questions that follow concern Open Access publishing. For the purposes of this survey, an article is Open Access if its final, peer-reviewed, version is published online by a journal and is free of charge to all users without restrictions on access or use." For more details see Dallmeier-Tiessen et al. (2011a and b, including supplementary materials) which includes a detailed documentation.

[58] See also, Simon Lambert for the SOAP consortium, http://www.slideshare.net/ProjectSoap/soap-symposiumtalkii [accessed September 12, 2012]

text analysis of the written responses to question 9 resembles the dominant positive response to question 9[59].

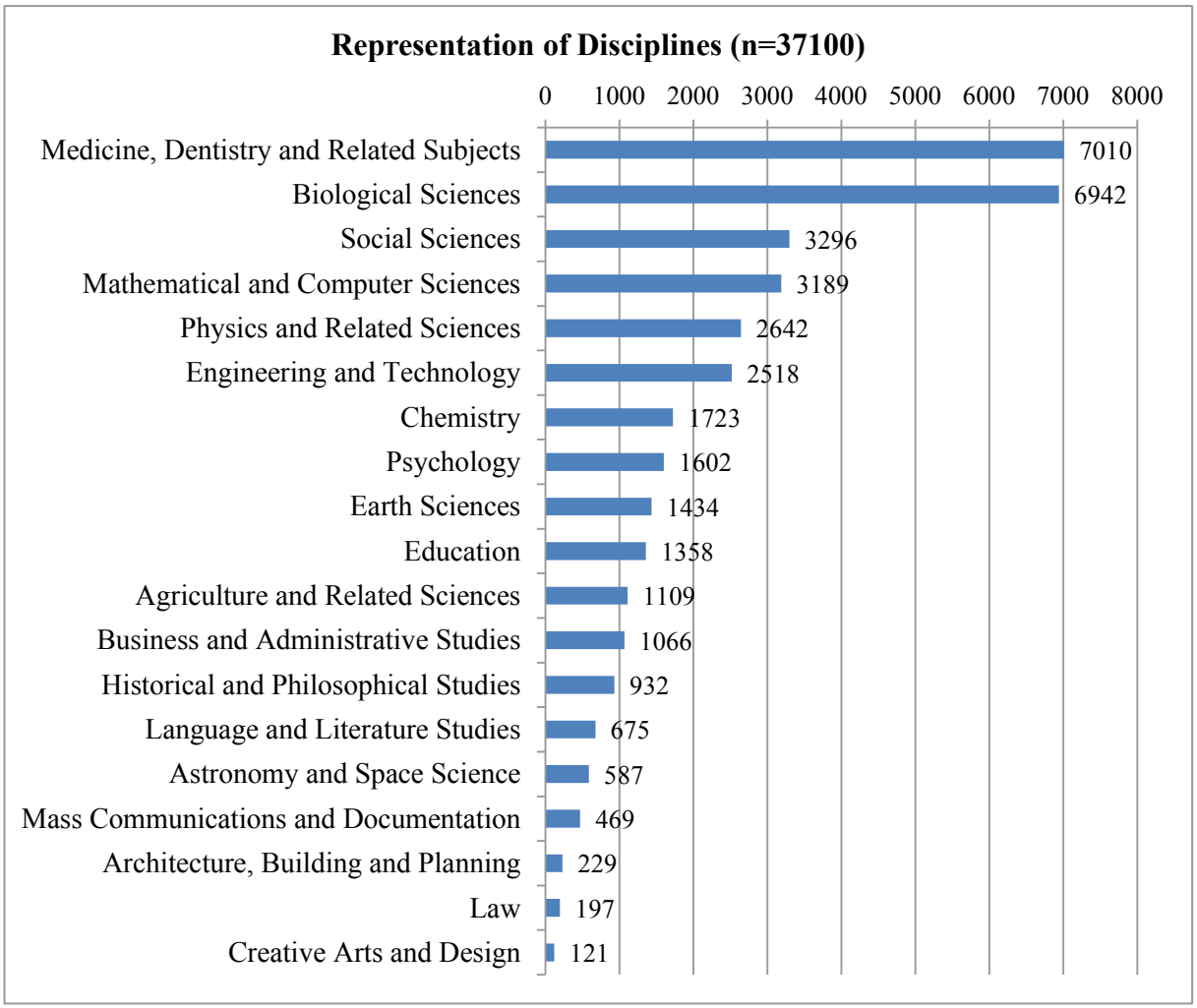
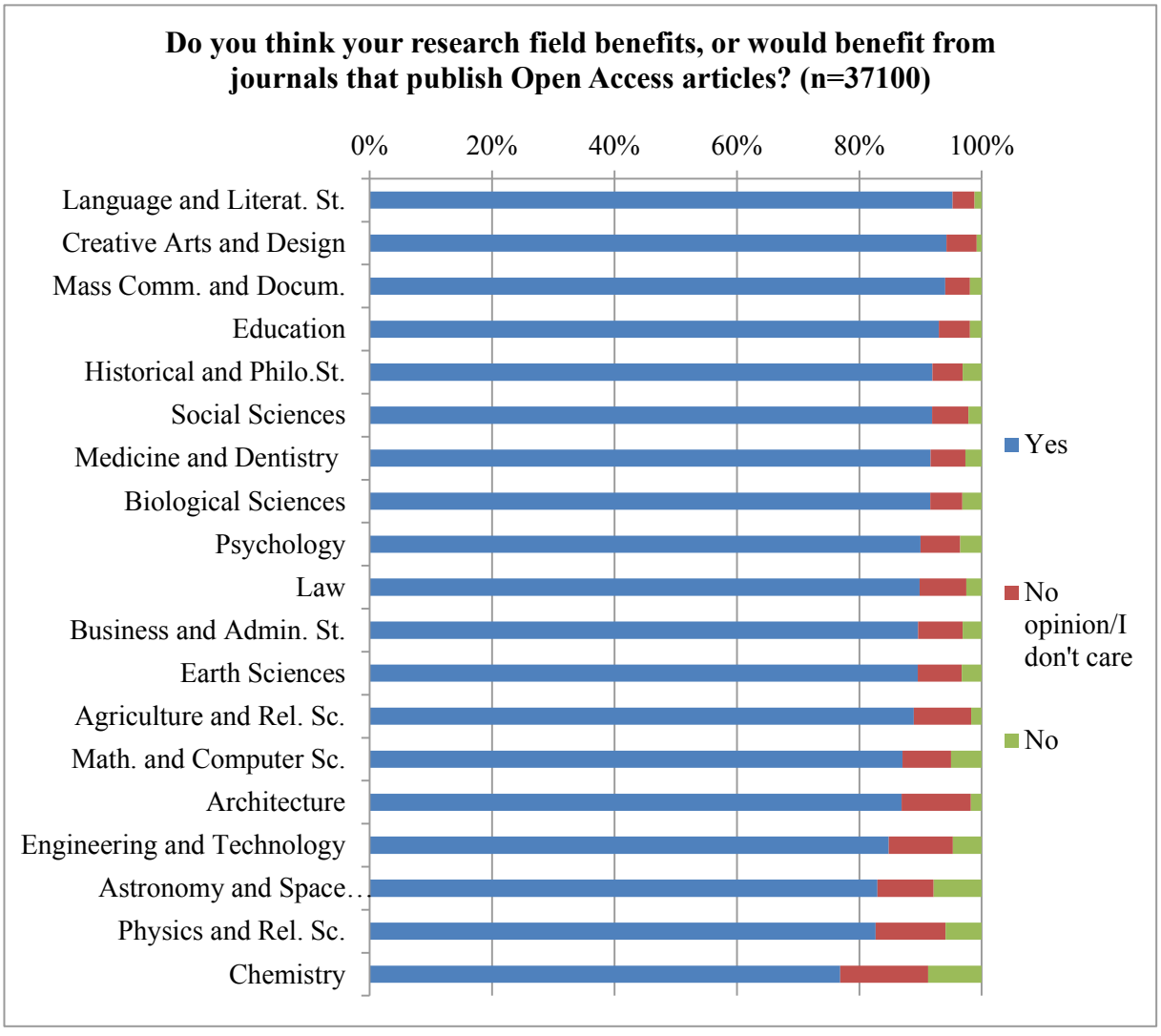


Figure 4: Distribution of responses to survey. Researchers' responses to the survey by primary research domain are shown. Only researchers who have already published at least one article (and who completed the main questions of the survey 9 and 2, see Dallmeier-Tiessen et al., 2011a) have been considered for this analysis. The results are presented in descending order.

[59] The response to question 9: the majority of the answers (22,312 tags) were considered "positive towards" OA, and only few were "negative" in their attitude (1,825 tags). This means that the respondents gave free text answers. The text corpus has been tagged for similar answers and topics. The resulting tags are grouped into positive OA views and more negative OA views. For more details see Dallmeier-Tiessen et al. (2011a and b, including supplementary materials) which includes a detailed documentation.

This mainly positive attitude towards OA publishing does not however reflect into concrete actions taken by researchers represented in the answers to question 15 ("Approximately how many Open Access articles have you published in the last five years?", Figure 6). On average 29% of the researchers stated that they had not published any OA article. 52% of the respondents had published at least one OA article and 9.5% more than 5 OA articles. Experience in the individual disciplines is very different, i.e. researchers in the biological sciences and medicine have more experience in OA publishing (more than 60% have published at least one OA article[60]). At the other extreme, 60% of the researchers in the astronomy and space sciences did not know about OA or had not published by OA means.



[60] It needs to be noted that these two disciplines have been addressed via the BioMed Central publishers in particular and thus it needs to be speculated that there is a bias towards OA publishing experience in these disciplines.

Figure 5: Question 9 of the SOAP survey. The question investigates the attitude of researchers towards OA publishing. The results are shown in descending order of the fraction of respondents replying "yes".

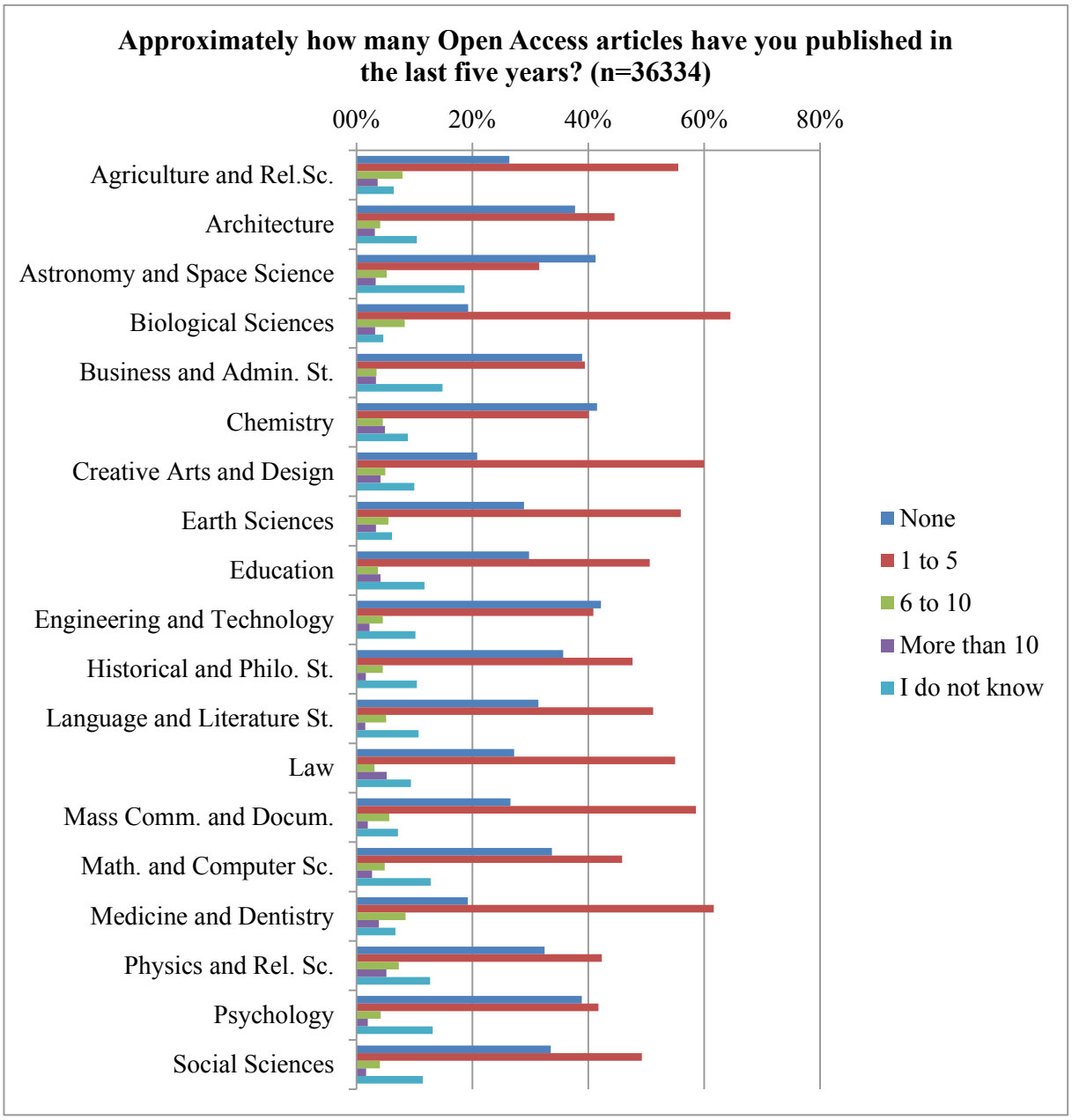


Figure 6: Responses to question 15 of the SOAP survey. The researchers' experience with OA publishing by disciplines is specific to the discipline. The results are presented in alphabetical order.

The pronounced difference between a positive attitude towards OA and publishing practice is further underlined by the response pattern to question 13. It asked researchers to rank factors relevant to their journal selection: the three most important factors for journal selection are "prestige/quality of the journal", "relevance of the journal for my community", and "journal impact factor"[61]. The factor "the journal is Open Access" is ranked as being less important or irrelevant[62]. This means, in 2010 OA is not highly relevant when selecting a journal for publication.

The responses to question 16 give a detailed idea of existing barriers to OA publishing ("Has there been a specific reason why you have not published an article by Open Access? If so, please give your reason(s) in the textbox provided", Figure 7)[63]. 42% of the researchers provided a written response. The analyzed free text answers show that funding and quality are perceived as the main drivers, 39% and 30% respectively. Accessibility, unawareness, and habits are issues, but mentioned less frequently. Researchers from different disciplines show very distinct responses. In particular in the natural sciences and medicine funding is by far the main reason not to publish OA. But many disciplines in the Humanities and Social Sciences (HSS) and also the earth sciences show a smaller fraction of the funding barrier to publish OA, i.e. social sciences (22%) and education (23%). The quality barrier also varies and is strongest in astronomy (49%), chemistry (39%) and business and administrative studies (37%). The survey comprised questions about the funding barrier, which revealed disciplinary and country-specific differences[64]. In question 19 researchers were asked "how easy is it to obtain funding if needed for OA publishing from your institution […]?" The majority of the researchers (54%) responded that it is difficult to obtain funds. 31% stated it is easy to obtain them; 15% did not use any. The differences in discipline and country suggest that solutions are at hand in some disciplines. One can speculate that this is linked to dedicated and tailored funding and support schemes on a national or disciplinary level.

[61] In this thesis the term „impact factor" refers to the measure of journals indexed in the Journal Citation Reports by Thomson Reuters. This is a measure based on citation counts, originally conceived by Eugene Garfield. http://thomsonreuters.com/products_services/science/free/essays/impact_factor/ [accessed August 17, 2012].

[62] 55.7% in total. See Appendix A for detailed figures.

[63] In addition, one could also study the response patterns to question 9 of the survey in more detail.. The free text analysis and tags to the question "Do you think your research field benefits or would benefit from journals that publish OA articles?" results in a list of detailed drivers and barriers. From the positive responses: Scientific community benefits (36%), financial issues (20%), public good (18%) are named most frequently (Dallmeier-Tiessen et al. (2011b), with data documentation for details about tags). There are only small differences related to disciplines. Studying the negative aspects raised, one finds two tags the most important: "low quality" and "not needed" (both 18%), see also, http://www.slideshare.net/ProjectSoap/soap-symposiumtalkii [accessed September 12, 2012].

[64] See also, Salvatore Mele for the SOAP consortium, http://www.slideshare.net/ProjectSoap/soap-symposiumtalkiii [accessed September 12, 2012].

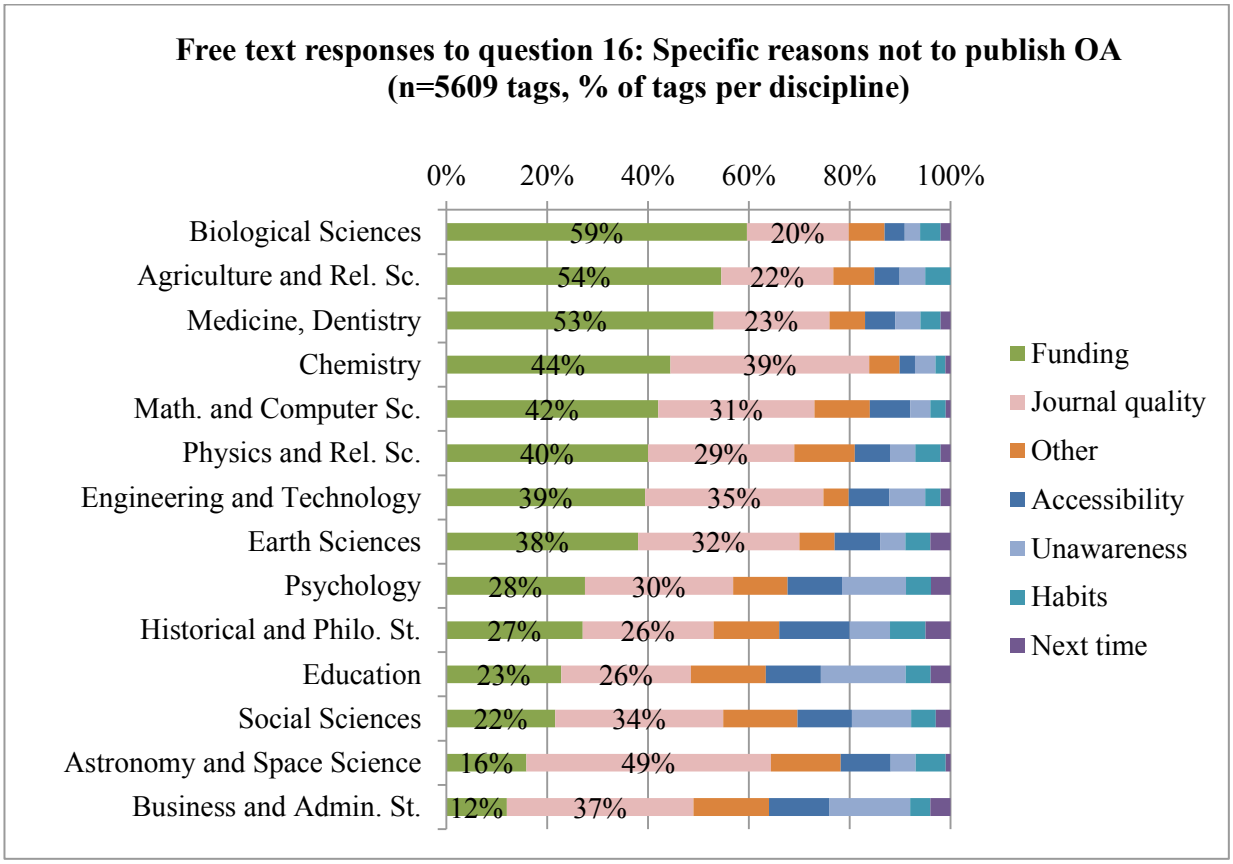


Figure 7: Barriers to publishing OA. Derived from the free text analysis to question 16 in the survey. Only disciplines with at least 100 tags are shown. The results are presented in descending order of the fraction funding.

In summary, two main barriers with regard to OA publishing can be highlighted: funding (costs) and quality. Moreover, it appears that quality and impact in scholarly communication are generally of major importance for publishing researchers – independent of an OA aspect.

These results have been supported by the results of the follow-up survey[65]. Here, 2,664 researchers replied to the question "Why did you publish OA?" They were able to choose multiple reasons, and selected "content freely available to readers", "quality/prestige", "quality/speed", "no fees" and "impact factor" the most often[66]. This points to some existing OA journals in selected disciplines.

[65] It should be noted that the preliminary results of the first survey were used to compile and phrase the questions of the follow up survey.

[66] See also presentation by Salvatore Mele for the SOAP consortium, http://www.slideshare.net/ProjectSoap/soap-symposiumtalkiii [accessed September 12, 2012].

Researchers who had not yet published by OA answered the question "what would make you choose to publish in an OA journal?" that "quality/prestige", "impact factor", "no fees/waiver" were the most important (Dallmeier-Tiessen et al., 2011c).

2.4 Discussion

The analysis provided the first comprehensive understanding of drivers and barriers in OA. In particular, they highlighted:

- The discrepancy between researchers' support for OA and their actual "publishing habits": the survey showed that 89% of the researchers across disciplines were in favor of OA, but only a small fraction of researchers had experience with OA publishing;
- The main drivers to OA publishing are: scientific community benefits[67], free distribution of content, quality and prestige[68], impact factor (if existing);
- The relevance of the main barriers to OA publishing: missing funding and quality & prestige.

The results allow for a new assessment of the current digital scholarly communication system. In the following, emphasis is given in the framework to the barriers of funding and quality & prestige, so that recommendations can be given to stakeholders for future work on these barriers.

The dominance of the funding barrier is interesting, as the SOAP study also revealed that the majority of the respondents had not paid any fees for their OA article in 2010 (Dallmeier-Tiessen et al., 2011b). This has also been seen by other studies (e.g. Shieber, 2009; Swan & Brown, 2004; Suber & Sutton, 2007). But, researchers perceive this as a strong barrier, probably related to new and (yet) unfamiliar business models such as article processing charges that often require the researchers to take action and organize funding[69]. However, the results presented here already indicate that there are strategies emerging to overcome the funding barrier, for example by corresponding and tailored funding schemes. This is evident in Germany, for example, where the DFG (German Research Foundation) supports institutional funding. Similarly, other examples in

[67] To allow a convenient reading flow the drivers and barriers are not given with quotation marks in this thesis.

[68] If existing for an OA journal.

[69] In opposition to the subscription based model, in which this is taken care of by the library.

the national or European layer exist, e.g. for projects under the umbrella of the 7th Framework Programme of the European Commission[70]. Also in the disciplinary layer support is available via collaborations with specific publishers (e.g. Copernicus Publications in the earth sciences[71]) or through membership schemes (e.g. for BioMed Central[72]). At the other end, initiatives are under way that target the amount of article processing charges. They highlight lower price ranges by keeping up standard peer review procedures for quality assurance. One example is PeerJ[73], an electronic journal, which advertizes article processing charges starting with 99Euros. Moreover, most OA publishers now have analogue models in place, fee waivers, etc. The results also indicate different kinds and degrees of access to funding in different disciplines, with, for example, a strong difference between the biological and medical sciences. Parallel studies (cf. Björk et al., 2010) have confirmed these differences in the disciplines. Results presented here confirm the different handling of fees in different disciplines, and also illustrated that some communities (e.g. earth sciences) find it easier to access funding than others (e.g. in the medical sciences). These results suggest that the barrier funding is surmountable with tailoured measures in place (cf. Van Noorden, 2012). Thus, more emphasis will be given to the other main barrier, prestige & quality.

The quality and prestige of a publishing outlet is a deciding factor for researchers when choosing a journal[74]. This is independent of OA first of all. This means OA is not at the forefront of aspects to consider in this process. It can be hypothesized that researchers take a perceived lack of prestige and quality of OA journals as a reason not to publish OA. These results have been studied further in the SOAP project via dedicated questions in the follow up survey[75]. The overall response pattern shows that researchers prefer to publish in established journals that have a high standing in the community. They are considered community approved and recommended by peers. Such journals are associated with the attributes prestige and quality[76].

[70] http://ec.europa.eu/research/science-society/index.cfm?fuseaction=public.topic&id=1300&lang=1 [accessed September 7, 2012].

[71] http://www.helmholtz.de/en/press/press_releases/artikel/artikeldetail/helmholtz_centres_facilitate_open_access_publishing_with_copernicus_publications/ [accessed September 8, 2012].

[72] http://www.biomedcentral.com/libraries/membership [accessed September 5, 2012].

[73] http://peerj.com/ [accessed August 1, 2012].

[74] See also results to question 13.

[75] See also presentation by Salvatore Mele for the SOAP consortium, Slide 36 http://www.slideshare.net/ProjectSoap/soap-symposiumtalkiii [accessed September 12, 2012].

[76] See also results to question 16.

The perceived lack of quality has been tackled in the communities in the meanwhile and examples emerge of surmounting this barrier. Generally speaking two strategies can be observed: the conversion of traditional (and community approved) journals to OA journals and the foundation of new ones.

For the first, this means that the business model and access model are changed from a subscription-based model to a new OA model (e.g. via article processing charges). Two examples stand out here: the journal Nucleic Acids Research[77] which moved to an OA model in 2005 and showcases a transition of a high-profile journal; and a large-scale initiative in the HEP physics community: the Sponsoring Consortium for Open Access Publishing in Particle Physics[78] (SCOAP[3]), which aims at rerouting subscription money for the main established journals of a whole discipline. This tackles the barriers of funding and quality & prestige at the same time.

The second strategy is the founding of a new high quality OA journal. Given the needed approval by the communities, such approaches might need a start up phase. But first successful examples exist. The OA journals from the Public Library of Science (PLOS[79]) have been assigned impact factors that place these journals in the top ranges in any community. Following the success of PLOS journals, similar examples emerge: for example, the new "eLIFE[80]" journal, cofounded by the Max Planck Society, the Wellcome Trust and the Howard Hughes Medical Institute.

It has been shown that disciplinary differences are apparent in the results. They have been seen in regard to both the funding and quality & prestige barriers. This is not surprising when considering the different disciplinary publishing cultures, availability of OA journals and funding options. In respect of quality and prestige, it has to be noted that some disciplines have a strong dedication to

[76] The relevance of quality in scholarly communication had been discussed and underlined, e.g. by Warlick & Vaughan (2007). The authors also put this discussion into the context of Open Access. They highlight that " free public availability and increased exposure may not be strong enough incentives for authors to choose Open Access over more traditional and respected subscription based publications, unless the quality issue is also addressed." It should be added that their study was conducted in 2007. In the meanwhile many more OA journals exist (see DOAJ development for example), many of them also with quality-assurance processes and even an impact factor assigned (cf. Björk & Solomon, 2012 and references therein).

[77] http://www.oxfordjournals.org/our_journals/nar/about.html [accessed August 8, 2012].

[78] http://scoap3.org/ [accessed September 5, 2012].

[79] http://www.plos.org/ [accessed September 5, 2012].

[80] http://www.elifesciences.org/ [accessed September 5, 2012].

journals that are assigned an impact factor (through the Journal Citation Reports[81]). Others, e.g. some disciplines in the HSS, focus on monographs. Of course, this impacts the researchers' perceptions of what is valued or not within the community.

This gives a first idea of the different response patterns, for example to question 16 („Was there a special reason why you have not published OA?") in the SOAP survey. Chemistry (39%) can be found on the top, with 20% in the biological sciences. Here, it can also be speculated that this correlates to the existence of "established" OA journals in the field. The results presented here point to an uneven spread of "community approved", "high quality" journals in the disciplines. It is known that some disciplines, such as biology, do have well established journals, for example PLOS Biology[82].

These examples indicate that wider awareness is emerging. But how widespread this awareness is needs to be investigated and followed up. In the most recent study Björk & Solomon (2012) investigated the quality and prestige aspect in more detail based on citation data[83]. They conclude that OA journals[84] are approaching the same scientific impact and quality as subscription journals. These results might point to a transition here. But further studies are needed to investigate disciplinary aspects and their correlation to current incentive and funding mechanisms in scholarly communication[85].

In summary, the results point to a gap between attitudes and practices with regard to OA[86]. The main barriers are funding, quality & prestige. Prestige is a reason to publish for researchers. It is

[81] http://thomsonreuters.com/products_services/science/science_products/a-z/journal_citation_reports/ [accessed September 6, 2012].

[82] http://www.plosbiology.org/home [accessed August 2, 2012].

[83] They find profound differences between "older" and "younger" journals, meaning that there are almost no differences in the impact received whether publishing in a "young" OA journal or a young subscription journal (disciplinary differences occur).

[84] OA journals that are indexed in the "Web of Science"(Björk & Solomon, 2012).

[85] The dataset of this study could serve as an evidence base for more detailed assessments on this theme.

[86] The discrepancy between the researchers' support for OA and their actual publishing practices has been seen in this study for the first time. Subsequent studies have confirmed these results. One year after this survey was launched, the European Commission (2012b) conducted a survey on "scientific information in the digital age" which yielded 1,140 responses, 37.6% of which were from researchers. The support of the research communities for OA was confirmed by the survey: 93.5% of the researchers agree that "publications resulting from publicly funded research should be as a matter of principle in the open access mode". The discrepancy with OA practices is also evident in the low share of OA articles worldwide. This can be seen in the studies by (Björk et al., 2010; Laakso et al., 2011). They report an overall availability of OA articles of 7.7% and 6.6% respectively. The SOAP study arrived at an availability of 9% of OA articles worldwide, which is in the same range (Dallmeier-Tiessen et al., 2010).

part of a researcher's visibility. A perceived lack of quality and prestige of OA journals is a strong barrier to OA publishing.

Overall, the results show that strategies are at hand to tackle the funding barrier. With the latest developments in the political layer across Europe (e.g. European Commission, 2012a), one can speculate that such enhanced support might increase in the future as well. This discussion has also shown that measures are underway to work on the aspects of quality and prestige as well. Many of them are long-term commitments and the results of journal conversion or the impact of new OA journals need to be followed up in the future as well.

However, it is necessary to investigate the connection of the two aspects to the current incentive system as well, and in particular to understand whether changes on that level are feasible. So far, such research assessments schemes are focused on publications and do not particularly assess OA publications. Even though policies increasingly demand sharing materials by OA means, such workflows are not incentivized in current research assessment schemes. In particular for a transition period on the way to more open sharing such strategies are worth considering. They can frame first actions to be worked on by the different stakeholder groups involved in digital scholarly communication. It is evident that the factors of quality and prestige are connected to the researchers' reputation, but more work on the connection to the current incentive system is needed. These aspects will also be considered in the case study in chapter 5.

2.5 The results within the framework of digital scholarly communication

The results presented in this chapter show that the significance of quality and prestige in a researcher's work style plays a central role in his decision-making in (digital) scholarly communication. This is also confirmed by the study (Procter & Williams, 2010) which states that "both as producers and consumers of information, researchers seek assurances of quality".

The funding barrier needs to be carefully considered in a more general discussion about digital scholarly communication. Gold OA is also a new business model in scholarly communication, meaning that by moving away from a subscription-based model monetary aspects are rerouted as

well. Furthermore, the survey provided evidence that even though this is perceived as a significant barrier, there are solutions to overcome this barrier on national and disciplinary levels[87].

The study presented here focuses on a particular aspect of digital scholarly communication, namely gold OA. A similar study to the SOAP study has been conducted with regard to green OA. The PEER[88] project also highlighted the central role of prestige in the choice of journal for publication, in particular for young authors at the beginning of their careers (Fry et al., 2011). The project underlined that in the eyes of a researcher career advancement and the impact of the journals in which they publish are linked. A smaller study on the green road to OA also confirmed the overall results discussed in this thesis. The survey run by the Repositories Support Project in the UK highlighted that researchers are strongly in favor of OA in general, and also of gold OA (69%) and green OA (80%)(Wickham, 2011). Similarly, they also mark a strong discrepancy between attitude and actual practices concerning OA. These results underline that there is a need for a strategy to implement OA recognition procedures in incentive and research assessment schemes. In this way the adoption of OA might be accelerated in the research communities.

[87] These are the two dimensions which have been studied in the survey; the influence of other factors needs to be studied further.

[88] http://www.peerproject.eu/ [accessed August 12, 2012].

3 Drivers and Barriers to Research Data Sharing

3.1 Introduction

Digital scholarly communication facilitates the sharing and integration of research objects beyond the traditional paper. Research data is not new - it has always been a primary product of the research workflow and in some disciplines was published in appendix to publications, as tables, for example[89]. In the meantime, with the pervasiveness of the internet, research and associated workflows have been transferred into the digital environment. This also affects the handling of research data which can now be shared and integrated into the publishing workflow more easily and more flexibly. There is now a demand by many stakeholders, funding bodies, society and policy-makers to provide access to research data over the long term, with due respect given to specific constraints, for example for sensitive data. The framework of long-term access to research data will ensure the integrity of research results and allow for reuse of materials in the future. In 2012, the demand for sharing research data has reached the highest political levels in Europe (European Commission, 2012a).

But so far it is the impression that sharing research data is not a common practice among researchers (Alsheikh-Ali et al., 2011; Nelson, 2009). It is also known that there are disciplines where data sharing is already an established routine, e.g. “genomics, systems biology, astronomy and crystallography” (Key Perspectives, 2010). But even in such disciplines, some hesitation in regard to opening up research materials can be observed (Campbell et al., 2002; Savage & Vickers, 2009).

Up to now there is no systematic and synoptic multi-stakeholder study that investigates why researchers do not share their data (openly). Few studies have been conducted with a focus on data sharing in the digital age, and there are even fewer studies that have investigated practices across disciplines.

With the results from the Parse.Insight project[90] (2007) a first survey was available that focused on data preservation (and only in a small part on data sharing) in selected research communities in the

[89] See also p. 13 (chapter 1.3.3) for more background on research data sharing, e.g. on the developments, stakeholders and policy makers.

[90] The Permanent Access to the Records of Science in Europe (Parse.Insight) project is funded in the Seventh Framework Programme, http://www.parse-insight.eu/ [accessed August 2, 2012].

natural sciences and in HSS (Parse.Insight Consortium, 2007; Kuipers & van der Hoeven, 2009). The respondents from the survey were mainly from Europe. Researchers participating in the 2007 survey are not eager to share their research data with others, "only 11% of the respondents make their data available for researchers within their research discipline" (ibid.). The authors note that the 58% of researchers making their data available "to researchers within their research collaborations and groups […] may not be [considered] a very high figure […]". In reference to possible drivers and barriers, researchers mainly perceive legal issues and possible misuse of data as problematic (both 41%), followed by "incompatible data types" and "lack of infrastructures" (28%), "lack of financial resources" (27%), fear of losing "scientific edge" (27%), "restricted access to data archive" (21%), "no problems foreseen" (16%) and other (10%).[91]

Tenopir et al. (2011) conducted a cross disciplinary survey in 2010. With the answers of 1,329 researchers they described a snapshot of practices and perceptions. An interesting response can be observed by the researchers: 74.9% said they "share their data with others", but only 36.3% state that others can access their data easily (counting "agree strongly" and "agree somewhat"). This looks very diverse on a discipline-specific level.[92]. They find that "barriers […] are deeply rooted in the practices and culture of the research process as well as the researchers themselves". The seniority of the researchers seems to play an important role. They find that senior researchers over 50 are more willing to share their data than younger researchers. In regard to possible barriers[93], 53.6% of the respondents state "insufficient time", 39.6% say it is the "lack of funding" and 24.1% say that "they do not have the rights to make data public"[94].

[91] Another survey focused on digital preservation has been published by Whyte et al. (2010) and personal communication. They conducted two surveys via the Digital Curation Centre (DCC, UK). Both surveyed DCC users and focused mainly on digital curation. Two questions deal with research data sharing. 40% of the respondents say that they "have […] made […] research data available for external users" (29% no, 13% don't know). In answer to Question 18, "what access issues does your project/unit face?" respondents cite "intellectual property rights" first (74% in 2006, 73% in 2009), second "access security/privilege issues" and "privacy/ethical issues", third "discovery and identification issues" followed by "insufficient technical infrastructures" and "data format incompatibilities". These numbers have to be taken with caution in regard to general applicability, as respondents necessarily had a close interaction with data preservation (as DCC users).

[92] 49% of the researchers in the atmospheric sciences, 43.9% of the biologists and 38.6% in the physical sciences state that others can access their data easily, compared to only 23.2% of the social scientists and 13% of the medical researchers. But these numbers have to be taken with a note of warning as the absolute numbers of researchers within a discipline is rather low (i.e. in medicine only 31 researchers) (Tenopir et al., 2011).

[93] Namely: reasons "for not making data electronically available".

[94] More answers are: "no place to put data" 23.5%, "lack of standards" 19.8%, just to name the first five reasons.

A qualitative approach was followed in 2007 by Swan & Brown (2008) with a focus on six specific disciplines and some cross-disciplinary aspects. The authors highlight discipline-specific characteristics and commonalities. It is underlined that sharing of "raw data" is relatively rare, but derived data is made available in many fields. Nevertheless it is stated that many datasets are not made "readily-accessible and re-usable". They report several drivers to publish data: among them altruism, encouragement from peers, and future collaboration. Barriers are: lack of career rewards and its non-representation in research assessments in the UK, lack of time, lack of expertise in data management and legal constraints. This report provides a UK focused overview[95].

Piwowar (2011) uses a bibliometric method to understand "factors associated with openly archiving raw research data", namely for raw gene expression microarray datasets. She concludes that "authors are most likely to share data if they had prior experience in sharing or reusing data, if their study was published in an OA journal or a journal with a relatively strong data sharing policy, or if the study was funded by a large number of grants by the National Institute of Health (NIH)". Authors of studies on cancer and human subjects were least likely to make their datasets available.[96]

The studies available give an overview of the past and current situation in sharing research data in digital scholarly communication. They validate anecdotal evidence (e.g. "Data's shameful neglect", 2009; Nelson, 2009) that research data sharing is not common practice and highlight different disciplinary practices. Furthermore, they also give a first idea of possible drivers and barriers on the researchers' side. This is also highlighted by Simon Hodson, who points to drivers and barriers but also reflects on the role of the stakeholders in the system to further data sharing in the communities (Hodson, 2009).

That's why this chapter focuses on multi-stakeholder approach, with special emphasis on the researchers. In order to improve the detailed understanding it is necessary to follow a qualitative approach. By using interviews particular topics of interest could be scrutinized more deeply. The perspectives of researchers are enhanced by viewpoints of different stakeholders involved in

[95] It appears (but is not evident from the report) that the study focuses on the situation in the UK and interviewees were recruited from the UK.

[96] More discipline-specific studies have been undertaken. Wicherts, Bakker, & Molenaar (2011) studied psychological papers and found that the reluctance to share data could be associated with weaker statistical evidence in a study. Sablonnière, Auger, Sabourin, & Newton (2012) focus on data sharing in the behavioral sciences. Wolkovich, Regetz, & O'Connor (2012) claim that data archiving and publishing in ecology is not done and propose a 3-step approach to solve this problem. Insights into different disciplinary practices are found in Key Perspectives (2010).

scholarly communication, such as data centre or repository managers, representatives of funding bodies, publishers, and others. These different perspectives help to elaborate problematic issues which otherwise often remain undeclared. Such insights would have been hard or impossible to capture in quantitative surveys. It is expected that the people working with the researchers within scholarly communication will further the understanding of drivers and barriers.

This is done via a two-step qualitative approach. A first round of interviews is used to identify drivers and barriers. A second round of interviews is used to peer review the drivers and barriers detected, to enrich and qualify them, e.g. by discussing their own expertise in that regard. Particular emphasis is given to experiences that included overcoming barriers.

The interviews were conducted as part of the European project "Opportunities Data Exchange"[97] (ODE) project, which is a consortium led by CERN. Project partners are CSC[98] (Finland), STM Assoc.[99] (The Netherlands), LIBER[100] (The Netherlands), Science and Technology Facilities Council[101] (UK), Helmholtz Association[102] (Germany), British Library[103] (UK) and German National Library[104] (Germany). The consortium consists of publishers, libraries, data producers and managers, research organizations and researchers, and thus provides access to all the relevant stakeholder groups dealing with data sharing and preservation. The project aims at "identifying, collating, interpreting and delivering evidence of emerging best practices in sharing, re-using, preserving and citing data." One focus of the work lies on the identification of drivers and barriers in data sharing. The author of this thesis contributed significantly to the project. The results presented here have been designed and composed for the thesis and are an independent achievement of the author of this thesis.

[97] ODE Project, funded for 2 years under the 7th Framework Programme by the European Commission, http://www.ode-project.eu [accessed July 22, 2012].

[98] http://www.csc.fi/english [accessed July 22, 2012].

[99] http://www.stm-assoc.org/ [accessed July 22, 2012].

[100] http://www.libereurope.eu/ [accessed July 22, 2012].

[101] http://www.stfc.ac.uk/ [accessed July 22, 2012].

[102] http://www.helmholtz.de/en/ [accessed July 22, 2012].

[103] http://www.bl.uk/ [accessed July 22, 2012].

[104] http://www.dnb.de/EN/Home/home_node.htm [accessed July 22, 2012].

3.2 Approach

In order to obtain qualitative insights into drivers and barriers in research data sharing, a two-step approach was developed based on expert interviews (Bogner, Littig, & Menz, 2009). In this way the broad subject could be narrowed down by a first round of interviews, and in a second step findings of the first round could be deepened and reviewed. Overarching themes could be scrutinized. In between the two rounds a workshop was conducted to review the first results and progress through external experts.

3.2.1 First round of interviews

The first round of interviews aimed to give a broad overview of drivers and barriers in data sharing from different perspectives, and so different stakeholders were interviewed: researchers, funders, and data managers. It was intended to conduct interviews from a wide range of disciplines, i.e. from the HSS and the natural sciences.

Setup: A semi-structured interview guideline and themes were developed, which would allow the experts to fully expand on their expertise and experience in regard to data sharing. A structured interview guideline at this stage did not appear feasible as it would have minimized any possibility to enhance the interview during its progress (Pickard, 2007). These first round experts were asked for their experience in data sharing, success stories and finally drivers and barriers (see interview protocol in Appendix B).

Data gathering: Project members suggested experts. It was intended to cover as many disciplines and roles as possible with the limited number of interviews. The interviews of the first round were conducted by four members of the ODE project consortium, seven by the author of this thesis[105]. In all 20 interviews were conducted in the first round. They were conducted over the course of several weeks during spring 2011 (see also Schäfer et al., 2011). Furthermore, these seven interviewees were chosen to showcase a diverse research and infrastructure landscape: HEP, molecular biology (genomics), HSS, economy, astronomy and, across disciplines, the perspective of a national funding body and the data repository perspective. The interviews were conducted on the telephone or video conferencing tools and took approximately 60 minutes.

[105] This refers to the interviews with participants 1 to 7.

Data analysis: A transcript of each interview was developed and approved by the interviewee. The interviews were shared in an internal evidence base for the project. The interviewers then provided the three to four main hypotheses from their interviews. The hypotheses were collected within the project; a set of categories was developed by the interviewers so that all project members could assign these categories to the hypotheses. In this way an overview of predominant drivers and barriers in the interviews was developed. The full analysis is presented in the project report (Dallmeier-Tiessen et al., 2011d; Schäfer et al., 2011).

The results from the first round of interviews are summarized in a list with brief explanations of drivers and barriers[106]. The list of drivers and barriers was peer reviewed by independent experts[107].

3.2.2 Second round of interviews

In the second round of interviews the revised set of drivers and barriers was to be reviewed and refined through additional expert interviews. Two interview guidelines were developed, one focusing on interviewees who are researchers and thus also possibly data producers, another one targeting the non-researchers[108].

Setup: For the second round all project members suggested experts in the field of data sharing[109]. The author of this thesis chose 19 interviewees according to their field of expertise and their role (researcher, publisher, infrastructure provider, or librarian)[110]. The number and selection allowed a comprehensive analysis of drivers and barriers across disciplines and stakeholder groups. Thus not only additional researchers, infrastructure providers and data centre managers were selected, but also several publishers, as they had not been present in the first round of interviews used in this thesis.

[106] This list is called "conceptual model" in the resulting ODE publications and in the workshop.

[107] This list was presented and discussed in the workshop which took place alongside the Alliance for Permanent Access Conference 2011 in London. The workshop was attended by project members and invited experts, including some from disciplines that were not represented in the interviews, e.g. clinical trials or biodiversity research. Their feedback was incorporated into the collection of drivers and barriers. The project members and interviewers had also drafted the new questionnaire for the second round which was presented. The interview structure was considered appropriate by the workshop attendants.

[108] Stakeholders who work with researchers, who provide services, infrastructures, policies, funding etc.

[109] Experts were detected through relevant publications, conferences, partnering projects, private/professional connections.

[110] 21 persons were interviewed in 19 interviews. Two interviews were conducted with two persons at the same time. The interviews were joined by interested colleagues. This concerns participants 12 and 25.

The semi-structured interview guideline comprised closed questions, multiple choice questions and open ended questions. It aimed at recording the personal experience of the interviewees and thus allowed the interviews to expand particular topics (see Appendix B for the detailed semi-structured interview guideline).

Data gathering: The second round of interviews was conducted starting at the end of January 2012 (see Dallmeier-Tiessen et al., 2012). The author of this thesis conducted 19 interviews[111], with 8 researchers, 4 publishers, 3 funders, and 5 data (preservation) managers (where multiple roles are possible, e.g. researchers being active in data preservation management). In addition to many cross-disciplinary views (e.g. by the funders and publishers), the following disciplines were covered or touched on: archaeology, clinical trials/medicine, crystallography, earth sciences/climatology, economics, linguistics, HEP, HSS, material sciences, and molecular biology. The list of drivers and barriers derived from the first round of interviews was sent to the interviewees in advance for preparation. The interviews generally took 45 minutes.

Data analysis: At first the author of this thesis identified themes in her own interviews by analyzing and tagging the interview corpus. These salient themes could have been topics or drivers and barriers that were missed or that needed to be qualified in the list of drivers and barriers. In addition, solutions or repeated examples of overcoming barriers were included. The relevant quotes for the given topic were compiled accordingly. This will be discussed in this thesis.

All interviewers of the ODE project (in total 5 persons) followed the same approach for their interviews, 55 in total. The interviewers discussed the progress of the interviews continuously. Finally, every interviewer presented the main "themes" that were dominant in the interviews they had conducted. The compiled collection of themes was discussed among the interviewers and a comprehensive list of themes was developed. Then, each of these overarching themes was studied across all the interviews[112]. The results of the second round of interviews are presented here solely focused on the interviews conducted by the author of this thesis. In a second step (chapter 3.3.3), the individual analysis is then discussed in the framework of the wider group results.

[111] Overall, 55 interviews were conducted by the project members of ODE in this round.

[112] This allowed the specific drivers and barriers to be investigated. More importantly, this step required the interviewers to go through all the other existing interviews to check for the specific theme they had been assigned. This workflow ensured that personal views, interview habits and personal interpretation were normalized through analysis and discussion in the group. The interviews had been stored in an internal evidence base of the project for internal use. Within this thesis they are handled anonymously.

3.3 Results

The results are twofold: first, the brief list of drivers and barriers that has been developed based on the first round of interviews is presented; and second, the main themes resulting from the full two-step course of the interview process and the validation of the conceptual model are discussed. For this analysis only the interviews conducted by the author of this thesis are used.

3.3.1 Drivers and Barriers (first round of interviews)

The first round of interviews revealed several hypotheses, that were tagged and grouped into the following drivers and barriers (for more details, see Schäfer et al., 2011; Dallmeier-Tiessen et al., 2011d).

Drivers

a) Societal Benefits

b) Academic Benefits

c) Research Benefits

d) Organizational Incentives

e) Individual Contributor Incentives

Barriers

f) Individual Contributor Barriers

g) Availability of a Sustainable Preservation Infrastructure

h) Trustworthiness of the Data, Data Usability, Pre-archive activities[113]

i) Data Discovery

j) Academic Defensiveness

[113] The term pre-archive activities refers to data preparation needed before data can be shared. This includes for example documentation that allows others to understand and reuse the data. For specific types of data, such as personal data for example, data confidentiality measures need to be undertaken.

k) Finance

l) Subject Anonymity and Personal Data Confidentiality

m) Legislation/Regulation

This list of drivers and barriers was discussed and endorsed by the experts in the workshop[114]. The drivers and barriers reveal a complex framework that influences an individual's or group's interest in research data sharing. Moreover, the results showed that the advancement of data sharing is multifaceted, and appears to be specific in particular disciplines. As one example the interview in molecular biology highlighted a data sharing culture that is very advanced. In response, they focus on different challenges (such as data deluge) in comparison to the other disciplines which have an impact on the drivers and barriers in the discipline. The results suggest that data sharing is not yet common practice in many parts of the HSS. The interviews also showed that there are innovative structures in place or emerging to tackle data challenges. Such aspects and in particular the interconnections need to be understood more deeply.

3.3.2 Outstanding themes in the interviews (second round)

The interviews of the second round also refined the list of drivers and barriers, but furthermore allowed the interviewees to further expand on their expertise and experience in data sharing. The text corpus was tagged and consolidated into several overarching themes. Some of them are of cross-disciplinary relevance; some are more specific to one or several disciplines. These themes are discussed in the following, solely based on the interviews conducted by the author of this thesis. Special emphasis is given to experiences that show how barriers are surmounted.

3.3.2.1 Cross-disciplinary theme: Culture of sharing and incentive system

Many interviewees mention different aspects of the social dimension of data sharing. It is often not explained further, but rather highlighted that there is a societal or psychological layer which goes beyond a technical layer. Many interviewees link it to the current incentive system in research which will be singled out towards the end of this theme. Given the context, in some cases one can presume that "culture" is used as a synonym for "tradition". Generally speaking this theme is

[114] More details on the workshop can be found in Appendix B.

perceived as a barrier, i.e. the lack of data sharing culture in many disciplines, but with the exception of some disciplines such as biomolecular sciences.

One can distinguish different layers:

- External factors, e.g. via funders, editors, journals, community (boards), influencing researcher(s) to share data,
- Interactions with(in) the local research group influencing researcher(s) to share data,
- Individual reasons that inhibit or convince researchers to share data.

It becomes evident that this theme is very discipline-specific: some report a data sharing culture, and some miss one. It is also to be noted that concerning a researcher's hesitation there is a difference in regard to "work in progress" and "finished" data that might for example be supplementary data to a published article. The fear of misuse – mentioned by many interviewees - can be incorporated into this theme. Nevertheless, all these layers include societal aspects and the quotes below underline this hypothesis.

Some interviewees mention the "hesitation" or missing data sharing culture and do not provide more details (participant 21 and 22, 2012). Another interviewee (data manager; participant 8, 2012) mentions the hesitation of a community as a whole and points out that this could be overcome with a strong collaboration of infrastructure providers with the community: "Hesitation in the community, […] in regard to sharing unfinished data is overcome by a strong collaboration with the community. This happens for example via […] projects where infrastructure and community are brought together."

The aspect of a community without a data sharing culture is also highlighted by another interviewee (data manager, researcher; participant 13, 2012) who states that "[t]here is just no culture in data sharing". The interviewee also highlights initiatives in the respective discipline focusing on technical aspects, but "they also discuss ethical issues and work on sharing awareness". He further reports that they would like to get "data archivists in place, which could be the key thing and a step towards establishing a data sharing culture."

One interviewee points out that the research environment plays a significant role in the decision process. It might be a consensus process rather than an individual making a decision. The interviewee (researcher; participant 11, 2012) states: "Within clinical cancer research there are usually smaller research groups who decide individually. In their competitive environment they are

rather hesitant to share data. This becomes even more evident when commercial partners participate in the projects. There is generally no data sharing culture yet and thus [it is] not very high on the agendas in the research projects."

One interviewee (data manager; participant 20, 2012) points particularly to a discrepancy between disciplines. He states: "Sharing across disciplines [means] we have to reduce cultural discrepancies as much as possible. The presentation of the data needs to be as neutral as possible so that everybody can understand and use them. This is a dream of course – as you need to add as much information as possible to enrich the cultural [discipline-specific] environment so that it becomes neutral information in regard to disciplines."

One interviewee (researcher; participant 25, 2012) highlights the decision making process of an individual researcher: "Data sharing or data provision is competing with paper writing on the priority list of a researcher – at the moment this is not a fair competition and that could be changed." The interviewee links this to the existing reward system in research and highlights the aspect of his personal attitude towards sharing. The interviewee shares as he "thought it was simply needed". In that regard he mentions aspects of overcoming barriers: "Barriers could be removed by integrating the data issue in graduate training schemes. That way one could work on the culture from the very beginning." The interviewee reports that "…in the biomolecular domain […] data sharing is well established".

One interviewee (publisher; participant 23, 2012) points to a closer work with the community to overcome barriers and remove hesitation to "…support researchers in preparing datasets, establish common data formats [and] efforts in regard to standardization." Another example is given by an interviewee who points to a framework of technological aspects and incentives (data manager; participant 22, 2012): "In regard to technology, there is also a lot that can be done. The field of astronomy has shown how tools like the virtual observatories can provide incentives to participate, and how issues like standardizations and formats can be overcome."

The topics of data sharing culture and the discussion of the current incentive and reward system in research are closely linked. Some interviewees point to the lack of rewards in the current system. One interview (data manager; participant 8, 2012) highlights that "There is anxiety to share, for example to steal "good ideas" without receiving benefit for the data production." Another interviewee (data manager; participant 22, 2012) highlights the missing reward as an important barrier.

One interviewee (funder, participant 12, 2012) explains that "it is difficult to convince researchers to share their data and it becomes obvious that policies are not enough." He further highlights that "it is not attractive to share - there is no reward for data sharing at the moment. In addition, many researchers are uncertain what to do".

Another interviewee (researcher; participant 18, 2012) explains in more detail that currently "peer visibility and status might be illusory if data is not curated properly. The only real reward might be self-re-discovery and re-use of the data in the future (if at all)." And another interviewee (researcher; participant 25[115], 2012) explains that "the most pressing issue is that published data is not a citable output often, the incentive system is focused on the paper which in [some] domains often requires the data to be shared." He explains further that this has an influence on the daily decision-making and prioritizing (see citation on previous page by participant 25).

3.3.2.2 Cross disciplinary theme: Financial aspects

The interviewees raise funding or financial aspects in general terms frequently during the interview process. In general terms, they distinguish two aspects. First the need for immediate and direct funding via projects, infrastructures and associated services, e.g. to establish corresponding repository infrastructures, work on metadata, interoperability challenges and data discoverability. Secondly the significance of sustainable funding is mentioned and discussed. Interviewees highlight the need of long term funding in particular to increase the researchers' trust to infrastructure, data and services.

This theme is – on a general level - perceived as a barrier. Solutions or examples of overcoming barriers are mentioned. Many interviews also link this theme to preservation. This could be a potential driver when sharing and reuse of material can reduce the spending of resources for further data production.

One interviewee (data manager; participant 20, 2012) highlights the overall costs of the research data lifecycle. He thinks this is "one of the most important aspects, it is an issue not to the single costs, but the "total cost of ownership" covering the whole lifecycle from the researchers with an idea, the funding, the research time, the report." The interviewee claims that "the researchers and most other people are not aware of the total cost of information that develops during the work, over

[115] This interviewee works in the field of molecular biology where data citation is already common practice and demanded, but was referring to the overall situation in the research system here.

the time of research. In the framework of publicly funded research this becomes even more complex as this data should be used and stored in a responsible manner etc."

Some interviewees focus on immediate investments and upfront funding and their links to data sharing. One interviewee (researcher, data manager, librarian; participant 13, 2012) states "some do want to share their data, but no budget is foreseen for sharing, for the effort, working hours, platforms, tools etc. There is a whole framework that needs to be changed."

One interviewee (researcher, participant 9, 2012) adds "data is expensive, software has to be there to process. There is indeed a lack of pre-archive and archive funding." But the interviewee has no worries about post-project questions. This statement is in opposition to concerns raised by other interviewees, who frequently discuss the mid and long-term perspective of data sharing and its connection to data preservation. One interviewee (researcher, data manager, participant 19, 2012) highlights that immediate costs might be low and could be covered by available funding right now, but asks "but what about the long term costs, e.g. covering the 10 years period?" Finally, one interviewee (publisher; participant 23, 2012) summarizes that "we need sustainable business models that allow an evolution of the data repository over time as well. Important question: who pays on mid [and] long term."

Some interviewees reflect on possible business models that exist or need to be developed (funder, researcher; participants 17 and 21, 2012). One interviewee states "It is also feasible to pay for data. If the data is of good quality that could work – if they receive funding for it. They receive money for computers, why should they not receive money for data [purchase]? So if they pay money for a computer, why should they not pay for data?" (funder; participant 21, 2012).

Another interviewee (funder, participant 21, 2012) thinks that "It is thus also needed to integrate data sharing in funding schemes and policy making." And thus points to solutions how this barrier can be surmounted. But funding and finance is not only perceived as a barrier, but also seen as a potential driver. One interviewee (funder; participant 12, 2012) highlighted that "reuse of materials could lead to cost cuts in data production." Another interviewee (data manager; participant 20, 2012) also mentions that it is possible "[…] to reduce costs of long term management by reengineering the cost model and data access."

3.3.2.3 Cross disciplinary theme: Infrastructures, standards and interoperability

The challenge of standards and interoperability is considered important in two ways:

- Standards and interoperability are needed to facilitate sharing (databases, data repositories). This happens on a discipline layer and across disciplines.
- Standards and interoperability are needed to enhance sharing and in particular reuse of data. This aspect touches on discoverability and services on top of the actual sharing process, e.g. in regard to the incentive system (see separate theme) – this aspect crosses disciplinary boundaries.

These aspects are perceived as barriers to research data sharing.

One interview, for example, highlights the need for common standards within a discipline. The interviewee (researcher; participant 18, 2012) states: "[I]n the field of crystallography they have agreed on common standards, such as the CIF[116] format. This means that such barriers have been removed. This also applies to Meteorology for example, where some standard values are easy to define, on a global scale. Thus, they have the advantage that they talk about the same things." Another interviewee (researcher; participant 25, 2012) points to a best practice discipline, molecular biology: "there are already discipline-specific guidelines which seem to work well. They also raise awareness and one need to see what is useful, what is not, what works, what does not."

Interviewees highlight the significance of data submission and ingest for this theme. One interviewee (data manager; participant 8, 2012) states that "the data repository takes a lot of burden from the researchers, [such as] metadata standardization and discoverability. A basic set of metadata is required upon submission of data." Another interviewee underlines the significance of researchers' support: he (publisher; participant 23, 2012) thinks that "community engagement – to support researchers in preparing datasets" is needed to overcome barriers in that regard.

In regard to the following steps in the research data lifecycle, interviewees underline additional challenges, e.g. preparing data for potential reuse. One interviewee (researcher; participant 18, 2012) highlights that "I would like to share my data, but to do it in a useful way you need to agree on common standards etc. Otherwise they are not useful for anyone."

[116] http://www.iucr.org/resources/cif [accessed September 12, 2012].

The aspect of discoverability, visibility and reuse is touched by other interviewees. One interviewee (publisher; participant 23, 2012) highlights the need for action: "[I]n regard to discoverability a lot can be done. Here, for publishers there is certainly a lot to do. They can get people engaged on their own platform where they provide an interface to access research data." Another interviewee (data manager; participant 8, 2012) points out that this aspect might become a driver for data sharing as well: "A well organized data repository benefits the researchers: it increases the visibility of the dataset which is a huge incentive." Thus, the interviewees connect this theme with the societal challenge and the incentive system.

Some interviewees specify some requirements, i.e. one interviewee (funders; participant 12, 2012) states "It is certainly important to permanently address research data. Good metadata is essential; context information is also needed, comprehensible metadata as well."

It has also been highlighted that there is an interoperability challenge for data sharing across disciplines. According to one interviewee (data manager; participant 22, 2012) this is due to "missing interoperable standards and formats across disciplines". One interviewee (publisher; participant 15, 2012) reasons that this is "[f]irst of all, because of the different data repository environments. Interoperability is an issue here. This makes discoverability an even bigger challenge. But it is important to highlight the need for good disciplinary repositories – there is no single repository or publisher. For sharing across discipline it is also important to have structured metadata on submission, also the data must be standardized and structured, as well as of good quality."

According to the interviewees this theme affects the researchers' attitudes and willingness to engage in data sharing. But interviewees also reported on initiatives which are working on these issues already, such as the […] directive in the geosciences (data manager; participant 20, 2012) or the DataCite[117] initiative (funders; participant 12. 2012).

3.3.2.4 Discipline-specific themes: Legal, economical or ethical constraints

Legal and ethical constraints in research data management and sharing appears to be a challenge mainly at a disciplinary level according to the interviewees.

[117] http://www.datacite.org [accessed September 9, 2012].

First of all, this is apparent in regard to studies focusing on human beings. Two interviewees (data manager, funder; participant 4 and 8) mention audiovisual material in linguistics and highlight the challenge of a corresponding legal framework.

Also in clinical research this theme becomes very relevant. One interviewee (researcher; participant 11, 2012) states that "[a] strong barrier in clinical trials data sharing is patient anonymity [with] personal data. […] [C]onfidentiality in regard to clinical trial data [is] very important". This is also underlined by another interviewee (researcher; participant 19, 2012) who reports a "strong issue with medical information. These data need to be anonymous. This is also important in regard to sharing – some records are not allowed to leave the UK." Another interviewee (publisher; participant 24, 2012) points out that "[o]ne should not be able to identify a person by the dataset."

Furthermore, commercial industries compete with research data sharing in some disciplines. One interviewee (researcher; participant 11, 2012) describes this in the field of clinical research: "It is important to note that there is a lot of data production and compilation for drug registration/application at a federal level. This means that such data is highly competitive, for example many industry partners are usually involved in conducting the research, and such data is then opened to these registration agencies, but not opened to the public. It is a very competitive environment."

One interviewee (researcher; participant 19, 2012) also uses external and commercial data in the geosciences. He highlights "for this research purpose we got a special agreement so that we can reuse the data for their models, but we are not allowed to share it any further, also not the derived data." This experience is shared by a researcher in economic research (participant 25, 2012).

Finally, one interviewee (data manager; participant 20, 2012) summarizes the situation as follows: "IP [Intellectual Property] is also a problem […] question of who owns the data. IP and security measures are a very complex field, also in regard to possible data policies. "

3.3.2.5 Discipline-specific themes: Archive activities and data preparation

Interviewees highlighted and exemplified discipline-specific practices and burdens in pre-archive activities. It has been highlighted that this activity varies greatly by discipline. This is in particular evident for the effort and time needed for data preparation. One interview (data manager; participant 8, 2012) points out that "Preservation is a main driver to submit data to this data

repository. The data repository has a very convenient setup, so that the data producer or submitter can tell the data repository which access restrictions apply. Reuse is controlled, licenses are used. […] the data repository takes a lot of burden from the researchers, e.g. metadata standardization and discoverability. A basic set of metadata is required upon submission of data."

One interviewee (researcher; participant 17, 2012) from the geosciences reports on important pre-archive activities that are needed to understand the data potential. He says that "There are two main barriers: time to support users and provide data to people who don't know the data. There is a lack of time and personnel. […] It is important to support these data re-users and to train them. They need to know the boundary conditions, e.g. IPCC[118] scenarios etc. That's why I provide this training and support."

Several interviewees note this challenge also in regard to medical data. One interviewee (researcher, data manager; participant 19, 2012) discusses "[a] particular example is for sure medical data. In the example of images in the neurosciences, these need to be processed and preserved in a way so that this is compliant with medical ethics." They found a solution to modify the framing of the images so that delicate (personal) information is "removed". His statement highlights that time and effort is needed for data preparation before sharing. This barrier needs to be discussed in the framework of the incentive system, mentioned before as a separate theme.

3.3.3 Results within the framework of all interviews conducted in the ODE project

After completion of the interview process the author of this thesis prepared the results (themes) independently. Afterwards, they were discussed within the group of project interviewers. The additional analysis including all interviews conducted in the project is published (Dallmeier-Tiessen et al., 2012). In comparison to the results presented here in chapter 3.3, the overall analysis based on 55 interviews allows for a more granular distinction of themes and discussions, but there are no significant differences from the results presented in chapter 3.3.2.

Themes that have been derived in addition by the analysis of the 55 interviews are:

- Role of publishers in data sharing,
- Data management: skills training and expert support,

[118] Intergovernmental Panel on Climate Change, http://www.ipcc.ch/ [accessed June 14, 2012].

- National and regional policy and legal frameworks,
- Public visibility of research data,
- Quality assurance of data.

The emergence of the theme "public visibility of research data" is particularly interesting in regard to the dominant themes presented in this thesis (chapter 3.3.2): missing data sharing culture, hesitation and incentive system. Legal frameworks have been mentioned, but policy frameworks have been partly neglected in this chapter: Even though they have been mentioned as part of community agreements (i.e. in the biomolecular science) in the interviews discussed in this thesis, the impact of top-down approaches has not emerged as a dominant topic within or across disciplines. This is especially interesting given the increased pressure by funding bodies and policy makers (cf. chapter 1).

3.4 Discussion

This study shows that the advancement of data sharing is very diverse and so are the relevant drivers and barriers. They differ very much with the discipline. It has been shown that data sharing is not widespread[119]; researchers do not share their data to a full extent. There are some prominent exceptions, e.g. in molecular biology, which is mentioned frequently as a pioneering discipline in data sharing. Interviewees also mentioned reasons for these advances: early community agreements of communities with journals, funders etc. which have been backed up by dedicated disciplinary data repositories and standards. But there are disciplines where data sharing is not yet a common practice, or is not considered at all in the research practices[120]. Some respondents reported that they share as much data as possible (within the legal or ethical constraints that exist); others report that such constraints make it impossible to share data within their domain. This is apparent, for example, in the interviewees that report from the HSS and medical sciences. According to the interviewees, there are signs of changes in the researchers' awareness, often due to the work of stakeholders (such as data curators or repository manager) who target communities at conferences, etc.

[119] Among the researchers who were interviewed.

[120] Most disciplines have been covered by the interviews. They have highlighted particular advancements in regard to data sharing, but they also highlight that it is difficult to generalize even within a discipline, as practices vary much according the individual research settings, institution or even research question.

The interviews did not only highlight different advances, but also point to very different practices in handling research data in the individual disciplines, which are reflected in the individual drivers and barriers. This is based on the particular workflow for data generation and the individual (disciplinary) characteristics of research data. The diversity of research data was remarked on frequently in the interviews: research data can be generated by an individual over years, or may be the result of big collaborations. It might comprise many smaller datasets or complex simulation frameworks. Materials that are affected by legal or ethical aspects need to be prepared specifically for open sharing, so that the respective rights are not infringed. Nevertheless, some interviewees reported on solutions for sharing data affected by legal or ethical constraints. One interviewee explained how modifications in medical images facilitated their sharing. But interviewees complain that this can be a time consuming process without compensation, and highlighted the significance of repository staff for such pre-archive activities. Data preparation for sharing is also complicated by different formats and standards that are being used even within a discipline. Finally, it has to be noted that some datasets might be relevant immediately for the rest of the community and society (e.g. medical data), whereas other data might be considered interesting after a given amount of time (e.g. earth science or archaeological data). Some interviewees who share data reported that they do it because they themselves have been looking for data to reuse and were struggling to find data (even though they knew it existed). They did not want others to go through similar time-consuming processes. The results suggest, however, that reuse of shared data is not common in many disciplines (e.g. in parts of the HSS), with the exception of disciplines such as archaeology, molecular biology[121] or earth sciences.

One interviewer stated, „there is a whole framework that needs to be changed“ (participant 13, 2012). It exemplifies that there is a strong need to further data sharing (and reuse). Targeted strategies are needed on disciplinary layers and on a cross-disciplinary layer. The results presented here highlight specific aspects that need special attention: technical & infrastructural, organizational & strategic, monetary, and societal. The results also show that these are strongly interconnected.

The researcher’s hesitation in respect of data sharing is evident and is considered a strong barrier. Linked to this theme, a lack of incentives has been singled out as a separate overarching theme.

[121] Here, within the Fort Lauderdale Principles for example – data is considered as a “community resource” product, in consequence of the Human Genome Project, which is a community resource project (Wellcome Trust, 2003). This means that the shared datasets together are the most useful and comprehensive source for further research.

Interviewees pointed out that data sharing competes with article publishing on the priority list. The extra effort that might be needed for data preparation, documentation and sharing is not compensated. Current research assessments have a focus on the text publication output (journal articles, monographs, conference proceedings, etc.). The timely and frequent output of articles is in many disciplines thus the most important action and thus reduces the time available for data sharing[122]. This time pressure even increases with the end of research projects and is furthermore aggravated through funding streams that have no budget allocated for data preservation and sharing during or after the project.

On that layer, with different research workflows also different influencing factors are apparent, and it is not always the individual researcher who takes the decision to share or not, but rather a group of researchers, a project or collaboration. And even if it is for an individual researcher to decide, he is likely to be influenced by community agreements and habits – the community-specific "culture of sharing". In addition, the existing and evolving policy framework puts pressure on the researcher.

These results highlight a missing link between data sharing and the incentive system that hinders the advancement and spread of data sharing. Researchers need to see a reward for the extra effort they undertake for data preparation. Thus, in the future, the current award and reward system needs to be considered in respect to data sharing and reuse[123]. Both should be reflected in adapted research assessment and incentive systems. With emerging initiatives like DataCite[124] and the Data Citation Index (Thomson Reuters, 2012), new metrics might open paths in that regard. This might kick off a research culture in which credit is given and exchanged with data sharing and reuse (see also "Credit where credit is overdue", 2009[125]). Such a research culture would have a direct effect on the individual's hesitation, and the initial barrier to share might be lowered. Furthermore, it is needed to investigate the impact of data sharing on the scholarly communication culture in general, as early studies have shown that data sharing is associated with an increase in citation which might be an additional driver for data sharing (e.g. Piwowar, Day & Fridsma, 2007).

[122] Swan & Brown (2004) and Tenopir et al. (2011) also highlighted "no time" as one of the main barriers. This also points to competition on the priority list of a busy researcher's life.

[123] Other features of digital scholarly communication, such as scientific blogs, might also need to be considered.

[124] http://www.datacite.org [accessed September 7, 2012].

[125] This topic has already been mentioned by Swan & Brown (2008), but the results here suggest that the topic has somewhat increased in its significance since then.

But it is not only the incentive system and research assessment frameworks that need to adapted; on the other side it is necessary to establish further workflows and practices that qualify data sharing for such frameworks. Data citation via persistent identifiers (DataCite) has already been mentioned in the interviews. This requires dedicated data repositories that facilitate such services. In addition, data journals (e.g. Earth System Science Data[126], GigaScience journal[127]) need attention that target quality-assured data documentation with data sharing in dedicated journals. In this way sharing is accompanied by a publication that is assessed in current research assessments. Furthermore, such an approach builds on the familiar article based publishing concept and thus could find easy community approval.

Furthermore, some interviewees had already pointed to the need to further engage the community in data sharing. It was reported that repository staff members had successfully engaged researchers who were hesitant. Such a consultancy role in data management might be an important task for information management in the future, and might be important for overcoming the hesitation of researchers when sharing data for the first time. As Piwowar (2011) states: people who have shared by OA means already are more likely to do it again.

The results of this study underline that substantial financial support is needed to establish and sustain data sharing over the long term. In particular, strategies to support data sharing, as it has been discussed here, need to be backed up by tailored funding schemes. Thus the emergence of more and more policies by funding bodies to preserve and share data will be accompanied by the development of specific funding schemes. As suggested in the interviews upfront investments to establish services that simplify and incentivize data sharing are needed. Such funding - and appropriate business models - are also needed over the long term to establish a trustworthy data repository environment.

Trustworthy data sharing environments are strongly connected to infrastructural aspects, interoperability and standards. It is perceived as an important, maybe even crucial theme, where profound improvements could help to steer a researcher's or community's attitude towards data sharing. Quality assurance procedures and certificates are measures to increase the trust in infrastructures and services. Even though initiatives are under way (e.g. with the Data Seal of

[126] http://www.earth-system-science-data.net/ [accessed September 7, 2012] and see also Chavan & Penev (2011), Dallmeier-Tiessen (2011).

[127] http://www.gigasciencejournal.com/ [accessed September 5, 2012].

Approval[128] or data journals as described above) such activities have not yet taken off. Furthermore, community standards and agreements beyond a discipline are considered important for facilitating usage and reuse of data. Interviewees highlighted the need for enhanced discoverability of shared data. Such efforts, in particular beyond a particular discipline, need better collaboration between the stakeholders involved in research data sharing. Interviewees pointed to two collaborations in particular – between researchers and data repositories, and between data repositories and publishers. This points to specific needs, in regard to services for data preparation and sharing in a repository, and, further, to connect data in repositories and text publications via publishers or repositories. This could mean enhancing and simplifying data sharing (interfaces and workflows)[129] by offering value-added services in repositories and enhanced discoverability services via publishers' portals.

A stronger collaboration of stakeholders is also needed to address the diversity of the policy and legal frameworks on national levels and on a global scale. Researchers as data producers, data users and repository/infrastructure providers need support in overcoming this barrier. Expertise in this area could be part of specific data management training for researchers, infrastructure providers and also for information management students[130]. Such training will be needed to further data sharing and the support provided to the research communities. According to some interviewees this is also important to raise awareness towards the opportunity of data sharing. It is a relatively new phenomenon and thus requires the adoption of new training schemes that tackle the individual challenges.

The results point to a strong societal challenge and strong hesitation in some research communities that need to be surmounted. The interviews underline the missing link from data sharing to the incentive system; the establishment of such needs the involvement of stakeholders like funders, policy makers, community representatives and infrastructure providers to build and establish corresponding services. Beyond the results discussed here, there is an agreement that collaboration with the community is needed to overcome barriers in regard to the societal layer. Friend (2010)

[128] http://www.datasealofapproval.org/ [accessed September 7, 2012].

[129] Further, Sansone et al. (2012) state that the research community "requires solutions for overcoming barriers that accommodate the current "wealth" of standards and resources, but hides it from users, thereby simplifying their efforts to meet (or ideally exceed) applicable reporting requirements." This statement also points to the need to simplify the data sharing process, i.e. it is needed to retain the full potential of research data and in parallel reducing the burden on the researcher' side.

[130] The interviews pointed to examples where library and information management staffs have been involved in best practices, e.g. the DataCite initiative. But no widespread involvement has been detected. It needs to understood better how this can be the case in the future.

points out that overcoming the barriers (scientific culture, funding, and incentives) may require widespread grassroots support. The results presented here show that this touches on infrastructural, technical, legal and policy aspects which need to be taken into account.

In summary, the interviewees' responses highlighted the fact that there is a set of reasons why they or others do not share their research data. The wide range of themes in chapter 3.3 underlines the different challenges that need to be addressed by different stakeholders involved in data sharing.

The results are further confirmed by a parallel quantitative study that was conducted between summer and the end of 2011[131] (European Commission, 2012b). The survey was open to everyone to participate; respondent were asked to rate "barriers to open access to data". The barriers "lack of funding for infrastructures" (80.8% of the respondents ranked it very important), "lack of incentives for researchers" (80.5%) and "insufficient national/regional strategies/policies" (79.2%) are at the forefront[132]. Overall, the results presented here give further precision to the results by Swan and Brown (2008) by highlighting disciplinary patterns, the interconnections and the complex framework with a strong societal layer.

The hesitation barrier as part of the societal layer is studied in more detail in the HEP case study. This focuses on research data sharing in HEP – a discipline in which data sharing is not yet widespread. Special emphasis is given to the potential role of information management in the engagement of researchers. The study also incorporates the implementation of data citation services via DataCite.

3.5 The results within the framework of digital scholarly communication

It needs to be discussed how transferable the results presented in this chapter are to the environment of digital scholarly communication in general. Even though research data is not a new object per se, in its digital representation new workflows and services now exist or emerge. The technical barriers that have been described are not only focused on this digital object, but can also be extended to the wider framework, i.e. code sharing (Ince, Hatton, & Graham-Cumming, 2012), protocol sharing, lab book sharing and the virtual research environment in general. In that regard,

[131] This study surveying 1,140 respondents by the European Commission was conducted during and after the period when the interviews analyzed here were conducted.

[132] The aspects "lack of mandates"," lack of data management" and "confidentiality" are ranked less important.

similar challenges can be found, in particular when dealing across disciplines. Interviewees touched on some of the issues associated with (programming or simulation) code sharing, for example, and also pointed to the missing link to the incentive system. In some disciplines, initiatives are under way to overcome such barriers, e.g. also via dedicated journals for such materials (e.g. Nature Protocols[133]).

Thus the societal layer with its focus on the incentive system is highly relevant for participation in any tool or service in digital scholarly communication (cf. also Nentwich, 2009). One of the interviewees highlighted that any action or participation is competing with publishing "traditional" articles as they are incorporated in the current incentive system by funding bodies, etc. Thus, with limited timeframes available, participation in any new tool might be neglected if not integrated into incentive (and funding) schemes (cf. Tenopir et al., 2011).

The funding barrier applies when considering the upfront investment of sharing facilities and (e.g. Science 2.0) services. But funding has also been highlighted concerning sustainable infrastructures, a topic that is closely linked to the issue of trustworthy infrastructures and data, according to some interviewees. This theme and its challenges might apply to many other tools and developments in digital scholarly communication. If funding and the persistence of a tool, workflow or service are secured over the long-term – would a researcher use it and share materials with it? This question links back to the incentive system and the researcher's hesitation, which appears to be the main challenge.

[133] http://www.nature.com/nprot/index.html [accessed September 12, 2012].

4 Summary of Drivers and Barriers

Two research tracks have studied two main challenges in digital scholarly communication across disciplines: OA publishing and research data sharing. They show that common challenges exist which can be transferred to digital scholarly communication in general.

The results highlighted strong disciplinary differences concerning the advancement of both activities. There are some pioneers who have widely adopted these innovations. Notwithstanding these advances and the pervasiveness of a digital research environment there is no widespread adoption of these innovations. This is somewhat surprising as the overall attitude to OA publishing and research data sharing is on the whole positive, but there is a profound gap between attitudes and actions.

Taking both research tracks (on OA publishing and research data sharing) together, several main themes stand out as drivers and barriers. Some are specific challenges varying significantly between disciplines, other themes are common across disciplines. This is evident, for example, when it comes to the publication habit and its link to reputation and current incentive schemes, which affect both, the experience with OA publishing and research data sharing. The themes are societal, technical & infrastructural layers, as well as funding & strategy, and are summarized below.

4.1 Societal layer

It is known that researchers publish in community-approved publication outlets, such as monographs, journals or conference proceedings. Their quality and prestige is fundamental when deciding where to publish. The results of the survey show that the perceived lack of quality and prestige of OA journals is a strong barrier for researchers to publish OA. Across disciplines, OA is considered beneficial for the research field, but this is not the most important deciding factor for publishing that way. Researchers prefer to publish with community approved workflows, depending on the discipline in monographs or peer reviewed journals for example. The publication in such outlets is seen as a contribution to the individual reputation. In particular in parts of the natural sciences, journals with an assigned impact factor[134] are considered to have a high

[134] There is a continuous discussion about the „impact factor“ and its usage and interpretation. It is a journal focused metric and does not focus on an individual article and its impact. Independent of the discussion, it is still an important indicator used today.

„standing“ and are preferred publication outlets. Thus, in some disciplines the lack of an impact factor, from which some new OA journals are affected[135], may be considered a barrier to OA publishing. It has to be noted that the latest studies[136] suggest that the differences in quality and prestige between subscription journals and OA journals are getting smaller.

The results of the OA study in this thesis also point to a missing link to the reward system, which currently does not incentivize OA publishing particularly. This missing link is also evident in the researchers' drivers and barriers to research data sharing: Interviewees mentioned that preparing data[137] for data sharing is competing with publishing papers on the priority list. During the interviews it was pointed out that the incentive system currently focuses on particular text publications and misses out on the research data domain.

This is evident, for example, in the Research Assessment Exercise (RAE) in the UK. The RAE demands four publications from a member of a department for a submission to the assessment. The academic job portal in the UK describes the situation as follows:

> *"As with the RAE (Research Assessment Exercise), if you have any strong publications (i.e. monographs or articles in world class journals) then you are more likely to be hired during this period because you will be able to offer something to your new department's submission."*
>
> *Quote from the main academic job portal in the UK[138]*

This results in caution about adopting new innovations, which has been observed by the interviewees. The theme "hesitation" is mainly dominant in regard to research data sharing and it is strongly connected to the researchers' reputation and the incentive system. Researchers hesitate to share their data - apart from the biomolecular domain or parts of the earth sciences; this has been reported as a barrier in several disciplines. The concept of hesitation refers to a missing culture of

[135] Many of the OA journals are relatively new and the Journal Citation Report's standards require a minimum time span of continuous publication of 2 years. See http://thomsonreuters.com/content/press_room/science/688332 [accessed September 20, 2012].

[136] Conducted in the meantime from 2010 to 2012. It has to be noted that there is a focus on natural sciences. The applicability to HSS is limited due to different publishing cultures and a focus on monograph publishing. But also there, initiatives are under way, e.g. wit Open Access Book projects.

[137] Data preparation comprises for example to prepare the documentation, provide enhanced provenance information, to issue release notes etc.

[138] http://www.jobs.ac.uk/careers-advice/working-in-higher-education/1561/the-ref-the-research-excellence-framework-2014/ [accessed August 12, 2012].

research data sharing. This implies that within many disciplines there is no tradition in sharing research materials beyond the text publication[139]. Thus, there are at most only a few community-approved workflows.

The term hesitation highlights the personal attitude and habit of the individual researcher, working group or community. Piwowar (20111) finds that people with experience in OA publishing are likely to do it again. So it is particularly relevant to discuss the initial barrier before sharing for the first time. It is relevant, when a researcher is asked to participate in a tool or to share data for the first time.

It has to be noted, however, that generalizations and comparisons between disciplines are difficult as the study shows that research, publishing and sharing practices vary tremendously. This applies to OA, but in particular to research data sharing. With the individual disciplinary characteristics of research data, the workflows and associated challenges also change. As one example, the case study in HEP will be used to understand this theme in more detail.

All stakeholder groups mention that research data sharing needs to be simplified and incentivized. Openness, in particular data sharing workflows and data citation practices should also be recognized in funding schemes. The statement regarding the RAE in the UK above exemplifies a dedication to „strong publications (i.e. monographs or articles in world class journals)“ in assessment schemes today. For the future, it is necessary to revisit today's assessment with a view to strategies that incentivize „openness“. OA and research data sharing are demanded by funding bodies and policy makers[140]. The results presented here so far indicate that there is a need to support both by reflecting them in the respective incentive systems.

[139] This also results in the lack of established quality assurance processes. Researchers do not want to be associated with "low quality" research outputs. Interviewees report a common fear of being associated with the wrong results and interpretations based on their shared data. Strategic data quality assurance workflows are not yet common. Such fears play a significant role in the decision-making process. A perceived lack of quality has also been seen in the research track on OA, most likely pointing to the fear to publish in a journal that is not „community approved“ or does not provide an appropriate community approved quality assurance.

[140] See Introduction (chapter 1) for details.

4.2 Funding and strategy

Both research tracks pointed to funding and highlighted the need for better strategies and collaboration in several aspects. The results indicate that a whole framework needs to be strategically addressed here, also comprising monetary, political and legal frameworks.

The funding theme has been dominant in both research tracks. First of all, researchers point to it as a barrier to OA. This is affected by new and different business models. The results from the survey and the discussion of emerging initiatives, however, suggest that strategies are in place to surmount this barrier concerning OA publishing and that different models exist. Some of these initiatives are of a disciplinary character (e.g. SCOAP[3]); other strategies operate on the national layer. The success of the initiatives needs to be observed in the long term to understand how best practices work and how analogous models might be of interest for other stakeholders.

In the interviews all stakeholder groups mentioned this theme as a barrier to research data sharing. Independent of the disciplines, interviewees highlighted the need for upfront to long-term investment. This is needed to build tailored services, sustainable und trustworthy sharing environments (infrastructures), and to train and pay data management staff. Trustworthy infrastructures were especially highlighted as requiring long-term business models. Dedicated funding streams for such infrastructures and staffing need to be designed and assessed.

But strategic approaches that work beyond disciplinary and national boundaries are not only needed for the funding framework. Challenges in the legal frameworks also need to be addressed collaboratively, e.g. to discuss appropriate global licensing models. Here, a leading example is given by the Knowledge Exchange Initiative[141].

Such collaborations are furthermore needed to build enhanced services, for example on the connection from paper to data. This offers the opportunity to support OA and research data together in a coordinated approach, particularly also in regard to technical interoperability. Such an extended view might cross-fertilize both tracks and further Open Science as a whole. All such collaborations need trained personnel. Interviewees highlighted the relevance of adequate training and education, possibly starting early in a career. In addition, in particular disciplines where legal and ethical constraints apply, the need for pre-archive data preparation support have been highlighted.

[141] http://www.knowledge-exchange.info/default.aspx?id=461 [accessed September 8, 2012].

This underlines that while disciplinary patterns exist, it is necessary for stakeholders to extend the view across disciplines, and beyond infrastructures, legal aspects etc. to further digital scholarly communication through training and support. An open sharing environment requires the adaptation of expertise and research frameworks, particularly in the transition period. This means, for example, that best practices from other disciplines (e.g. a successful OA journal, data repository or data citation services) might be well of interest to other communities and need to be exchanged.

4.3 Technical and infrastructural layer

The lack of interoperability between systems is a strong barrier that has been highlighted in the interviews. This also points to the link between publications and research data, but focuses in particular on research data sharing, data repositories and associated services[142].

The lack of standards, agreed community formats and documentation are seen as a strong barrier by many interviewees, and is considered to be more difficult across disciplines. The aim of research data sharing is to allow further reuse of materials as well as reproducibility of research results. In that respect, it is essential to support trustworthy repository environments and services that facilitate long-term preservation and reuse. Such services need to reflect the required link to the incentive system. Moreover, enhanced visibility and discoverability to further reuse are needed across disciplines (while coping with disciplinary standards or retrieval methods).

Interoperability challenges within the infrastructure services in digital scholarly communication hinder sharing and restrict possible reuse scenarios of research data. Overall, there is a perceived need to address this theme as a prerequisite to the widespread sharing of research data. It is particularly interesting to remark that this applies to every discipline – even though the disciplinary workflows are remarkably different. But the need for progress is demanded in data-intensive disciplines as well as in disciplines with smaller datasets.

[142] If we extend the view to the green road to OA, this discussion needs to comprise a note on the technical framework of repositories for text publications as well. This has not been part of this study, but studies suggest that interoperable and standardized approaches are needed here as well to further submission to repositories and to enhance discoverability and value-added services. This also applies, of course, to the connection of repositories for text and data and to value added services that link repositories and incentive systems.

In summary, the results show that practices in digital scholarly communication are not fully exploiting the potential of Open Science. Notwithstanding some disciplines which are at the forefront in exploring such workflows, for many researchers „open“ publishing of research materials is not yet a common practice. This study shows that specific steps have to be undertaken to achieve the vision of a „global and interactive representation of human knowledge, including cultural heritage and the guarantee of worldwide access“, as stated in the Berlin Declaration (2003).

It has been shown that the incentive system is a strong driver for any action in digital scholarly communication. It focuses strongly on the number, prestige and quality of (text) publications. Thus, researchers hesitate to invest time and effort into new publishing models (e.g. OA publishing) or into research data sharing. It has been shown that many other themes are interconnected, but currently the societal layer appears to be the dominant challenge in digital scholarly communication.

Two particular aspects of the societal layer will be studied and refined in practice: the aspects of quality and prestige as part of a researchers reputation, and hesitation. In the following case study in HEP, they will be reviewed with special emphasis on the link to the incentive system. In this way, the relevance of such themes within an individual research community can be studied. Furthermore, this study within a research community will help understanding a potential role of information science and management to support such communities in surmounting barriers and engaging with new tools in digital scholarly communication.

5 Case study in High Energy Physics: Understanding Drivers and Barriers in Practice

5.1 Introduction

The first research chapters highlighted a researcher's reputation and his hesitation when dealing with new scholarly innovations as prominent themes[143]. Both have been chosen for this disciplinary case study. This practical approach facilitates a better understanding of the detailed disciplinary characteristics and of the relevance within the community. It is furthermore hypothesized that these two themes allow a significant involvement of information management.

The High Energy Physics (HEP) community will be used for this case study. The author of this thesis had been embedded in the community engagement strategy of INSPIRE (at CERN), while conducting the research presented in chapter 2 and 3. Building on the close interaction with the HEP community and the emerging results from the two research tracks, the idea arose to study the results in practice, i.e. study if the barrier could be surmounted. Tailored services have been chosen for the study that would allow "hands-on" testing. The design and fulfillment of the research presented in this chapter have been done by the author of this thesis.

First, this chapter highlights HEP-specific characteristics in regard to drivers and barriers using the results from chapter 2 and 3, desk research and the work with the community[144]. In a second step, the two (independent) case studies in HEP are described: the set up, methodology, results and discussion. The first case study focuses on the researchers' reputation. It studies the researchers' engagement with a new tool in digital scholarly communication. The second one focuses on the hesitation barrier in research data sharing. Both case studies will consider the relevance of the link to the current incentive system (as it has been seen in the previous chapters).

[143] The discussion has shown that there are solutions at hand for example for the theme funding (chapter 2), furthermore it is not feasible to test it in such a case study. Other themes (e.g. the technical one) are touched indirectly in this case study, e.g. through a collaboration between two community platforms (INSPIRE and arXiv) and also as part of the second case study that focuses on research data sharing in HEP. Also legal aspects that appeared as a barrier are touched in the second case study for example (licencing).

[144] Since December 2009 the author of this thesis has been working closely with the community at CERN, in particular with the digital library INSPIRE.

The applicability to other disciplines will be discussed, which could help supporting researchers in digital scholarly communication beyond the HEP community (chapter 5.7). Finally, the case study will be used to understand the influence of information management on HEP's engagement (chapter 5.8) with new tools and opportunities in digital scholarly communication.

5.2 The High Energy Physics Community

The HEP community is quite small. It consists of about 30,000 active researchers plus many who come from adjacent disciplines, such as nuclear physics, accelerator physics, cosmology, particle physics or astrophysics (Igo-Kemenes et al., 2010; Gentil-Beccot, Mele, & Brooks, 2009). The High Energy Physics community is generally split into two major groups: theoretical physicists and experimental physicists[145].

This distinction is particularly important with respect to scholarly communication, as both groups can be described as rather different in their research workflow and publishing habits. Theoretical physicists usually work individually or in smaller groups, while experimental physicists work in international collaborations of sometimes more than 3000 researchers, resulting in publications by more than 3000 authors (Igo-Kemenes et al., 2010; Weiler, 2012). Due to the size and complexity of these experiments, these researchers also often work together locally, for example at laboratories like CERN.

The HEP community offers the unique opportunity of an instrumental case study. This is mainly because it can be considered as being tightly knit and easily accessible for the author of the thesis to carry out the research. This is mainly due to the centralized research infrastructure provided by laboratories such as CERN, as well as to a tradition of centralized disciplinary information platforms that offer easy access to the community.

HEP can be described as a discipline with specific publishing habits (Gentil-Beccot, Mele, & Brooks, 2009). It is often referred to as the founder of the green road to OA: researchers drove a change to a strong parallel preprint sharing culture as early as the beginning of the 1950s (Goldschmidt-Clermont, 1965; Heuer, Holtkamp and Mele, 2008), mainly due to the long processing and production time in the established journals in the field. The preprint server arXiv, developed 1991 by Paul Ginsparg (cf. Ginsparg, 2011), is still today the major cornerstone in HEP

[145] More distinctions are of course possible, i.e. more granular distinctions beyond the two groups presented.

preprint exchange (Gentil-Beccot, Mele, & Brooks, 2009). Nevertheless, publishing research in HEP is still driven by submission to the high-profile journals of the community.

Today, preprints and the final journal articles are available via the HEP-specific digital library, INSPIRE[146]. Preprints are available in fulltext as the content is ingested from the repository arXiv; for final journal articles only metadata are available mainly. INSPIRE is the successor to SPIRES, a bibliographic database developed with the emerging preprint culture, first based on record cards, then moved into the online era. SPIRES had been serving the community for decades (since 1969), in terms of searches for citations and publication lists. Within the community there is a strong prevalence in the use of such community-specific systems such as SPIRES and its successor INSPIRE as well as arXiv in comparison to general search engines such as Google scholar (Gentil-Beccot et al., 2009).

The digital library INSPIRE replaced the platform SPIRES in October 2011. With approximately two searches per second, it is the main working tool of the HEP community in scholarly communication. Today, INSPIRE is jointly operated by CERN, Deutsches Elektronen-Synchotron (DESY[147]), Fermilab[148] and SLAC National Accelerator Laboratories [149]. It collaborates closely with HEP publishers and other information providers in HEP such as arXiv, the NASA Astrophysics Data System [150] and the Particle Data Group[151]. In total it comprises one million records with half a million OA full-text documents. Furthermore, INSPIRE comprises additional databases and directories. Among these, HEPNames[152] is a directory of researchers in the HEP and adjacent disciplines. Due to manual curation for decades, the directory is considered rather complete and thus comprises the email addresses and details of most HEP researchers[153]. The INSPIRE digital library also comprises a blog[154] and a Twitter stream[155] which both aim at

[146] http://www.inspirehep.net [accessed August 1, 2012].

[147] http://desy.de [accessed July 8, 2012].

[148] Fermi National Accelerator (Fermilab): http://www.fnal.gov/ [accessed July 4, 2012].

[149] http://www.slac.stanford.edu/ [accessed July 4, 2012].

[150] The SAO/NASA Astrophysics Data System: http://adswww.harvard.edu/ [accessed July 8, 2012].

[151] http://pdg.lbl.gov/ [accessed July 24, 2012].

[152] http://inspirehep.net/collection/HepNames [accessed July 8, 2012].

[153] It is an important feature for HEP researchers who submit information voluntarily and use it to locate former colleagues.

[154] http://blog.inspirehep.net/ [accessed September 20, 2012].

[155] https://twitter.com/inspirehep [accessed September 20, 2012].

communicating new services or features (e.g. new or better search queries) and raising awareness towards temporary operational errors in INSPIRE. Users are furthermore invited to inquire or give feedback via email; this feedback system is ticket based and preserved.

The user, so the researchers, are in the focus of the study which was conducted by Gentil-Beccot et al. (2009). The survey asked, among other things (see Figure 8), if researchers were willing to participate in "tagging" publications on a platform. Interestingly, 63% are willing to spend 30 minutes a week on this task. At the time of the start of this thesis, the actual willingness to participate in such tools had not been tested.

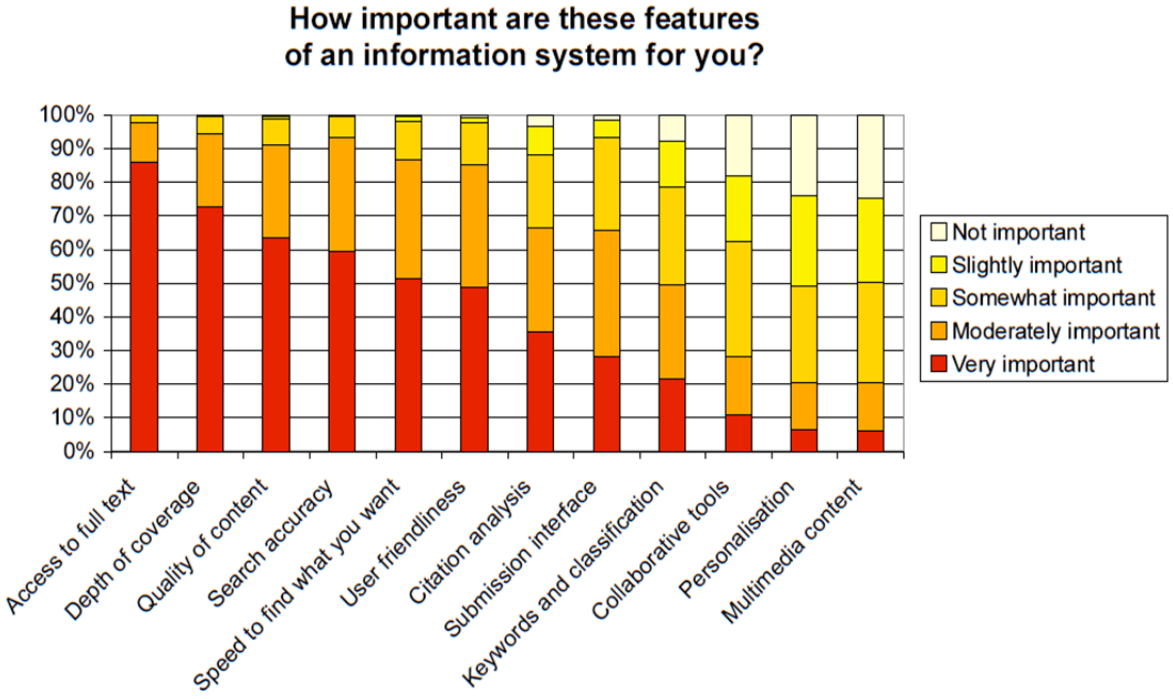


Figure 8: Significance of specific features of an information system for HEP researchers. Figure taken from (Gentil-Beccot et al., 2009). The results provide an overview of possible drivers for an engagement with an information system, e.g. via enhanced services that allow full text access and better discoverability ("search accuracy" above). The results also show that quality is a topic here, focused on the content of an information system.

Within the past decade the information systems in HEP have been facing new challenges, one of them the emerging need to deal with research data. Within the community HEP awareness has grown in regard to mid- and long-term data preservation[156]. The initial focus was on digital

[156] When this thesis was started; see also interview with Peter-Igo Kemenes, see Dallmeier-Tiessen et al. (2011d).

preservation, but not on the wider framework of integrating this into scholarly communication (e.g. access models, linking materials). With the Study Group for Data Preservation and Long-Term Analysis in High Energy Physics (DPHEP[157]) a first step was undertaken to change this in 2009. Data preservation in HEP is a specific challenge, as the data output is considered very complex and specific, even within the experiments (Brumfiel, 2011). Proposals were also made to use INSPIRE as a tool for data preservation.

In conclusion the reasons for choosing HEP (and INSPIRE as the technical component) as a case study are:

- The unique opportunity to have easy access to almost a whole community via email (HEPNames directory);
- A digital library that is established as the main working tool for the HEP community (INSPIRE);
- Availability of a 'one stop shop" digital library that is open for services and tools to be deployed (INSPIRE),
    - In terms of user engagement;
    - In regard to data preservation.

The applicability of the results of this case study to other disciplines will be discussed at the end of this chapter. As the frameworks and conditions in other disciplines look different this will be of importance. The case study was carried out by the author of this thesis at CERN, where she is employed as part of the Scientific Information Service team.

5.3 HEP-specific aspects of drivers and barriers

Two themes have been selected for the case study: reputation and hesitation. This chapter highlights the discipline-specific aspects of these drivers and barriers. This is based on the work with the community (e.g. from the current usage and feedback from HEP information platforms), literature review (in particular the results from the Parse.Insight project) and the results from chapters 2 and 3.

[157] https://www.dphep.org/ [accessed September 11, 2012].

Researchers can contact the digital library INSPIRE via email to ask questions, recommend features or demand a correction in specific records for example. Weiler (2012) noted that 49% of the researchers who contacted INSPIRE did so to inquire about their citations, while another 31% contacted INSPIRE to correct author-related data. The visibility of researchers to their peers appears to be of significant relevance; their profile needs to be correct and coincides with the stated significance of the researchers' reputation. This is also evident in the daily routine of INSPIRE, where normal instabilities in the number of citations are followed by numerous user requests. INSPIRE blog posts which inform about citations are the ones most frequently visited or read[158].

In regard to data sharing, more insights can be gained by consulting the Parse.Insight survey and the results from the interviews carried out in the ODE project (chapter 3). The prevalence of the topic of visibility and reputation has also been shown here. The Parse.Insight project conducted a survey among HEP researchers. The survey aimed at understanding the status of data preservation in HEP in 2007. The survey highlighted the relevance of the researcher's reputation in HEP in the framework of data preservation and data sharing, i.e. showing a strong fear of possible misuse of data which could impact the reputation (Holzner, Igo-Kemenes & Mele, 2009). The HEP-specific results from the interview analysis (chapter 3) showed that "sociological aspects" play a significant role in data sharing in HEP[159] (Dallmeier-Tiessen et al., 2011d).

In the work with the community the topic of data sharing has been present from the beginning of this thesis (beginning of 2010). It became immediately evident that conversations about data sharing were dominated by discussions about preservation, technical challenges and the fear of misuse (e.g. somebody might publish wrong results and this reflects poorly on the experiment). Openness was not a topic that was being discussed. These recent discussions thereby confirmed the currency of the Parse.Insight results (obtained in 2007; cf. Parse.Insight Consortium, 2007).

In summary, the HEP-specific aspects in regard to the two themes of hesitation and reputation (prestige, quality, incentives) highlight the strong significance of publications and citations within the HEP community. The visibility and correctness (quality) of a researcher's profile are of the highest importance. Also, there is a strong hesitation in research data sharing, with a prominent fear of misuse.

[158] One post on corrections in citation statistics received more than 1,300 hits within two weeks. An average blog post receives 300 hits over two weeks.

[159] Funding plays a role as well: "Barriers, or sort of barriers: some do want to share their data, but no budget is foreseen for sharing, for the effort, working hours, platforms, tools etc. There is a whole framework that needs to be changed" was the comment of one interviewee (researcher, participant 13, 2011).

5.4 First case study: The "reputation" driver in the HEP community

5.4.1 Introduction

This chapter comprises the first case study in HEP. It focuses on the reputation driver. In practice this driver was used to engage researchers in a new crowdsourced[160] tool and workflow on a new information platform. A tailored engagement strategy was developed to elicit participation in a new tool. It is hypothesized that a better knowledge about drivers and barriers helps in engaging a community. The HEP community served as the study environment. The community-specific information platform INSPIRE was used for implementing the framework needed for the study[161].

A specific online service for author disambiguation was chosen for this case study. The challenge of author disambiguation emerged due to the international and complex working environment, in which for example several authors with the same name appear on the same paper[162]. This required the development of a new service on INSPIRE. In order to be able to display correct author pages with the individual publication and citation details these name ambiguities needed to be solved.

Thus, at CERN an algorithm for author disambiguation was developed[163] that integrated a crowdsourcing approach. The author of this thesis designed and conducted the research presented in this chapter, namely the crowdsourcing experiment[164]. Researchers are asked to verify the output of the algorithm via dedicated interfaces. Researchers are approached by email to invite them to participate in this new service, to visit this interface and to manage their publications on INSPIRE. The workflow and user engagement strategy was defined and designed with a focus on the

[160] The term „crowdsourcing" is defined by Holley (2010): „Crowdsourcing uses social engagement techniques to help a group of people achieve a shared, usually significant, and large goal by working collaboratively together as a group". For this case study it needs to be highlighted, that the tool is open for everyone for usage, but it is intended to be used by a specific user community, the HEP researchers. The overall goal is to disambiguate the scholarly content in HEP and adjacent discipline so that all researchers have correct publication and citation lists and statistics.

[161] When this test was started, the community was still used to INSPIRE's predecessor SPIRES, which did not provide any enhanced participatory Web 2.0 services or similar features.

[162] For more details, see Brooks et al. (2011) and Weiler (2012).

[163] There are more author disambiguation workflows and tools available, for example Authorclaim by Thomas Krichel (http://www.authorclaim.org, accessed September 22, 2012). This chapter focuses on the study of the driver reputation and crowdsourcing. The algorithms and workflows in other disambiguation tools are rather different in their technical setup; and a technical comparison is not the purpose of this chapter. Thus for more details it shall be referred to Brooks et al., 2011 and Weiler, 2012. A comparison to the researcher's participation to other disambiguation services is given in the discussion in chapter 5.4.4.

[164] The corresponding author disambiguation algorithm has been developed by Weiler (2012).

reputation driver and its specific aspects in HEP. The user engagement strategy involves sending emails to researchers based on their entry in the HEP directory (HEPNames). But the interface is accessible for everyone to use.

The following chapter describes the approach in more detail. Then it points to the quantitative analysis of the user engagement, before discussing the results in terms of influencing factors and its wider impact.

5.4.2 Approach

This part of the case study in the HEP community aims at studying the reputation driver in more detail. This is done with the implementation of an engagement strategy. The respective workflow is implemented on INSPIRE, the digital library in HEP. The results are evaluated quantitatively.

The users of INSPIRE are presented with a workflow that allows them to verify, reject or add their publications via their personal publication list that are based on the clusters (the outcome of the algorithm; Brooks et al., 2011). They can do so as guests or with their arXiv credentials[165]. By collaborating with arXiv, usability and the researchers' trust in this service are increased.

The email text (Figure 9) invites the researchers to participate by highlighting significant terms for HEP researchers. On the basis of the drivers and barriers analysis previously described (chapter 5.3), these pointed to reputation-related terms in HEP, i.e. the publications and citations record:

- Highlight "publication list" in the subject line.
- Mention the connection between SPIRES and INSPIRE to improve trust into this new service and platform. The term "citesummary[166]" is used as it has been a popular feature on the previous SPIRES platform. This should also improve and increase trust in the quality of the new workflow.
- Mention "publications" and "citations" several times in the email – as these are significant drivers for HEP researchers.

[165] Researchers need an approved account when submitting publications to arXiv. This can be used to participate in the service.

[166] http://inspirehep.net/help/citation-metrics [accessed September 11, 2012].

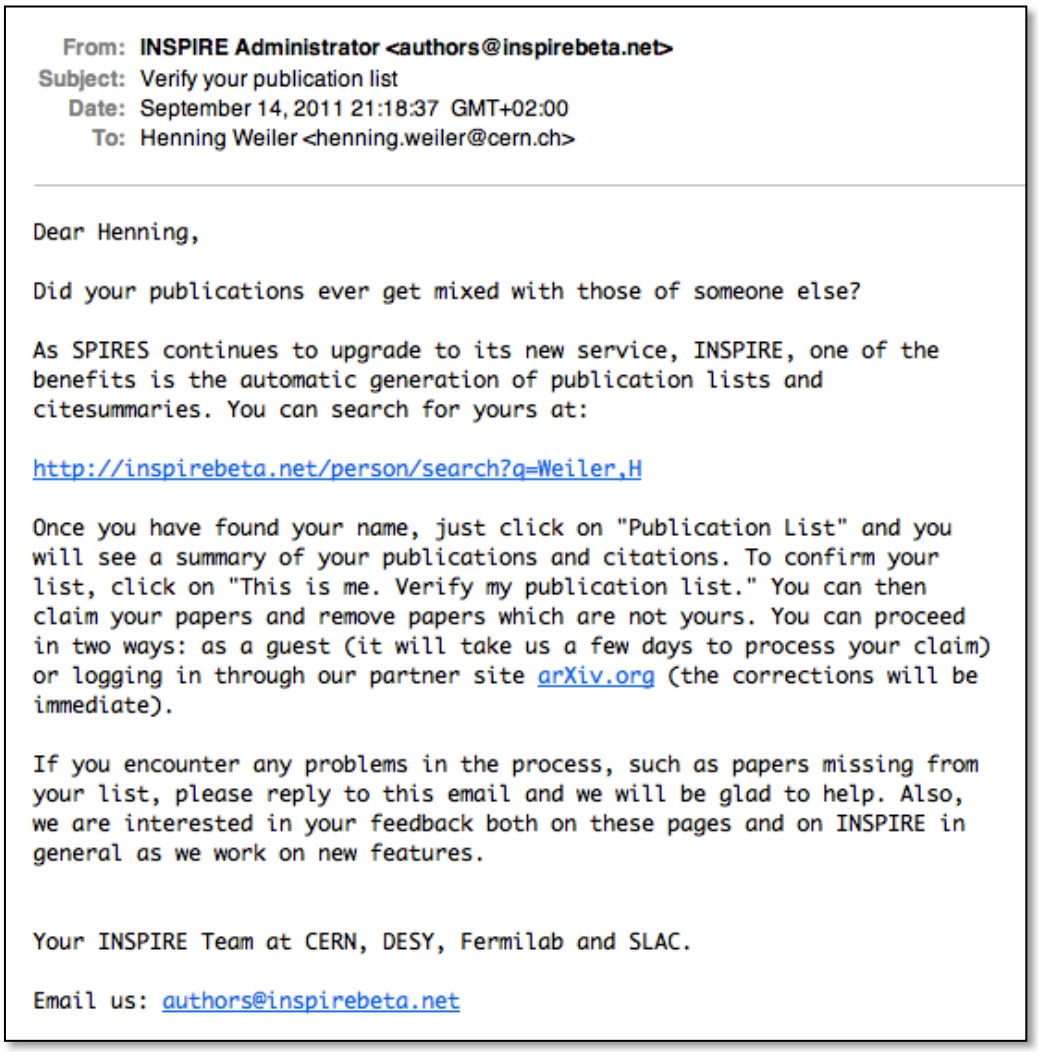
- Mention a collaboration with arXiv (which is the main OA platform in HEP) to increase trust[167] in the quality of the services and platform.
- Sign the email with the four main laboratories to increase trust in the service provider and make a distinction from any "spamming" agent explicit.

Within the email there is a direct link to the INSPIRE search[168]. Researchers who click on the link reach a prefilled search box with their name. This is possible due to the integration of information available from the HEP directory HEPNames.

Once researchers click on their name's link, they reach their respective author page. These popular sites on INSPIRE display an author´s publications and citations (Figure 10). This could happen, for example, when they search for themselves on INSPIRE. Users can also find the workflow and these author pages by serendipitous discovery or once it is recommended by other users.

[167] Trust emerged as an aspect in the cross-disciplinary and HEP-specific drivers and barriers analysis as well; for this reason it is used in this test as well.

[168] First there was a link to Inspirebeta, then to INSPIREHEP from October 2011 onwards.



From: **INSPIRE Administrator <authors@inspirebeta.net>**
Subject: Verify your publication list
Date: September 14, 2011 21:18:37 GMT+02:00
To: Henning Weiler <henning.weiler@cern.ch>

Dear Henning,

Did your publications ever get mixed with those of someone else?

As SPIRES continues to upgrade to its new service, INSPIRE, one of the benefits is the automatic generation of publication lists and citesummaries. You can search for yours at:

http://inspirebeta.net/person/search?q=Weiler,H

Once you have found your name, just click on "Publication List" and you will see a summary of your publications and citations. To confirm your list, click on "This is me. Verify my publication list." You can then claim your papers and remove papers which are not yours. You can proceed in two ways: as a guest (it will take us a few days to process your claim) or logging in through our partner site arXiv.org (the corrections will be immediate).

If you encounter any problems in the process, such as papers missing from your list, please reply to this email and we will be glad to help. Also, we are interested in your feedback both on these pages and on INSPIRE in general as we work on new features.

Your INSPIRE Team at CERN, DESY, Fermilab and SLAC.

Email us: authors@inspirebeta.net

Figure 9: Email invitation: The email text used to engage researchers in a new tool (on INSPIRE which is also a relatively new platform for HEP researchers who are used to the predecessor SPIRES) is shown. The mailing very much focuses on two aspects related to the driver "reputation" in HEP, as identified in previous research: publications and citations. The emails were first sent out by "Inspirebeta" and also linked to "Inspirebeta"; later these were replaced by "Inspirehep" links and emails.

In the following the workflow is described using John Ellis (theoretical physicist at CERN) as an example. He reaches the author page (Figure 10) and clicks on the link "This is me. Verify my publication list". He can choose between different publication profiles (which correspond to the clusters identified by the algorithm). Once he has chosen one, he is asked to either proceed as a guest or log in via an arXiv account.

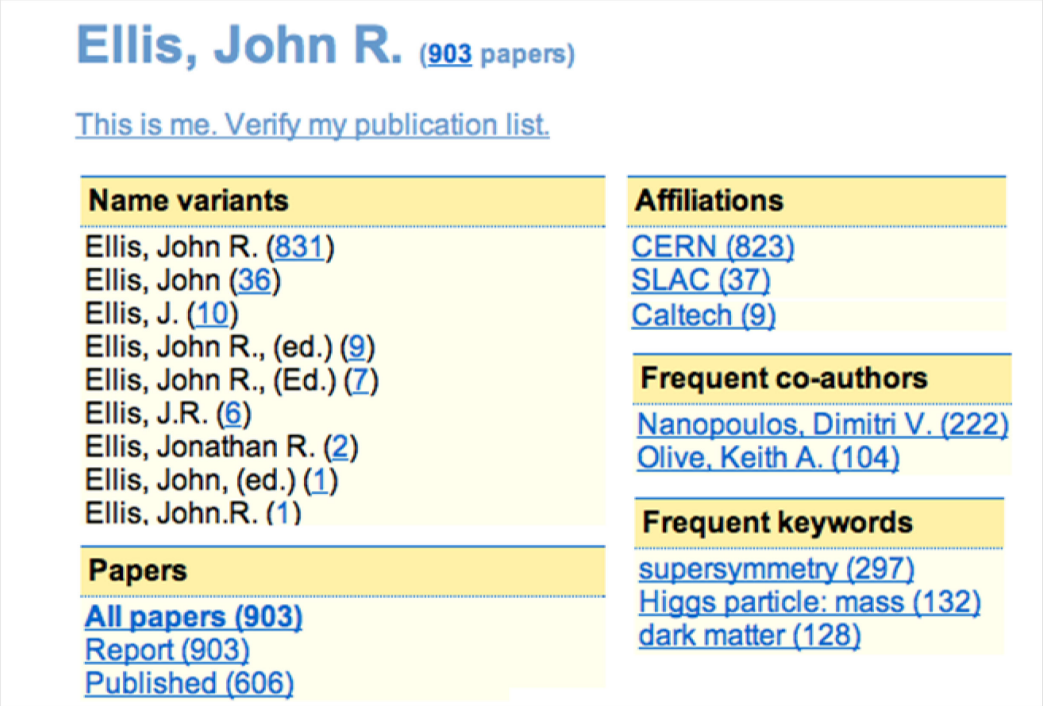


Figure 10: Screenshot showing the INSPIRE author page of John Ellis. This page is seen by researchers when they click on the link in the email and follow the profile selection menu (Figure 9). As of June 2012 this has been replaced by a new design integrating more databases.

He then sees a list of publications assigned to this publication profile. He can proceed and confirm his authorship, reject the paper or assign it to another person (Figure 11). He confirms the actions and receives a confirmation via email. He is also notified that changes submitted as a guest have to undergo an approval process by an INSPIRE editor. Changes submitted via arXiv are transmitted immediately.

In this case study user participation is studied to understand if a community can be engaged in a new tool (by using the knowledge about drivers). As part of the engagement strategy, groups of researchers are selected from the HEPNames database[169]. The email invitations are sent out by the INSPIRE team at CERN and Fermilab starting from early summer to the end of the year 2011. This is done in batches and by time zones so that emails are received by Tuesday lunchtime in the respective time zones. This should increase the participation in the tool, as it is known to be a suitable time for such international mailing campaigns.

[169] The selection was based on the latest update to the record. The records which are up to date are used first.

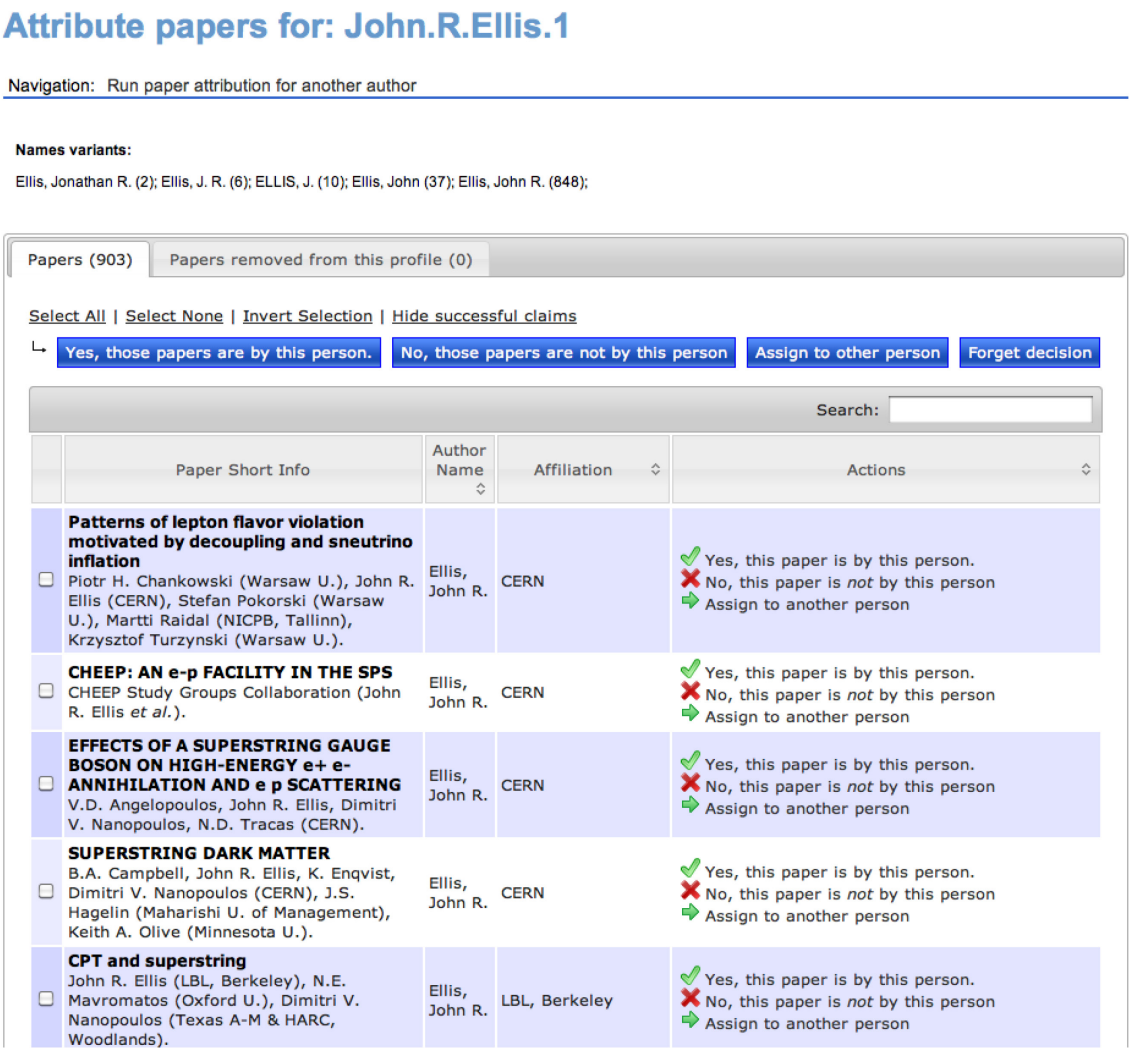


Figure 11: Interface for researchers to claim their publications. This action includes confirmation, rejection or assignment to another person. Researchers can log in via an arXiv account or proceed as a guest; information is then either transmitted directly to the database or is approved by an editor.

At first, mainly theoretical physicists who usually have an arXiv account (and thus could use the dedicated workflow) and whose publication lists are less complex[170] were addressed. Their details had been extracted from the HEPNames directory. The HEPNames directory has been curated manually over decades by the SPIRES and INSPIRE editors. This is important to note as the number of emails that bounced back is negligible, given that the email addresses are updated regularly in the database.

[170] In comparison to the experimental physicists.

Researchers were addressed in email chunks of up to a couple of hundred researchers. For this test, we identified the number of researchers approached by email and the number of researchers who replied. Beyond that, more specific details such as the workflow they used (guest or arXiv workflow) and the timing of the response (within hours or days) were recorded. Also the activity of researchers who did not receive an invitation, who presumably found the interface by serendipitous discovery was recorded. A high participation in the tool would be seen as success of the engagement strategy, built on the reputation driver.

In addition, the number of processed scholarly artifacts was evaluated. This was to give further insights in regard to the opportunities this approach affords for information management of information platforms and digital libraries. The analysis took place in the second half of 2011; the data snapshot was taken on November 28th, 2011. The dataset thus comprises data gathered over 38 weeks. The system is in production (on INSPIRE).

5.4.3 Results from the first case study

In the time span of the case study 2,558 researchers treated their publications via the interface. A total of 164,811 claimed artifacts have been recorded (see Figure 12). For the first weeks no user engagement (i.e. targeted mail-outs) was done, but also at this early stage frequent serendipitous discoveries were recorded (Fig. 12).

The first user engagement was started soon after the launch of the tool (week 24). The impact is visible in the cumulative plot due to a sharper increase (Figure 12). The response rate to the individual mail-outs is on average 40%. Some mailings reach a maximum response rate of 50%.

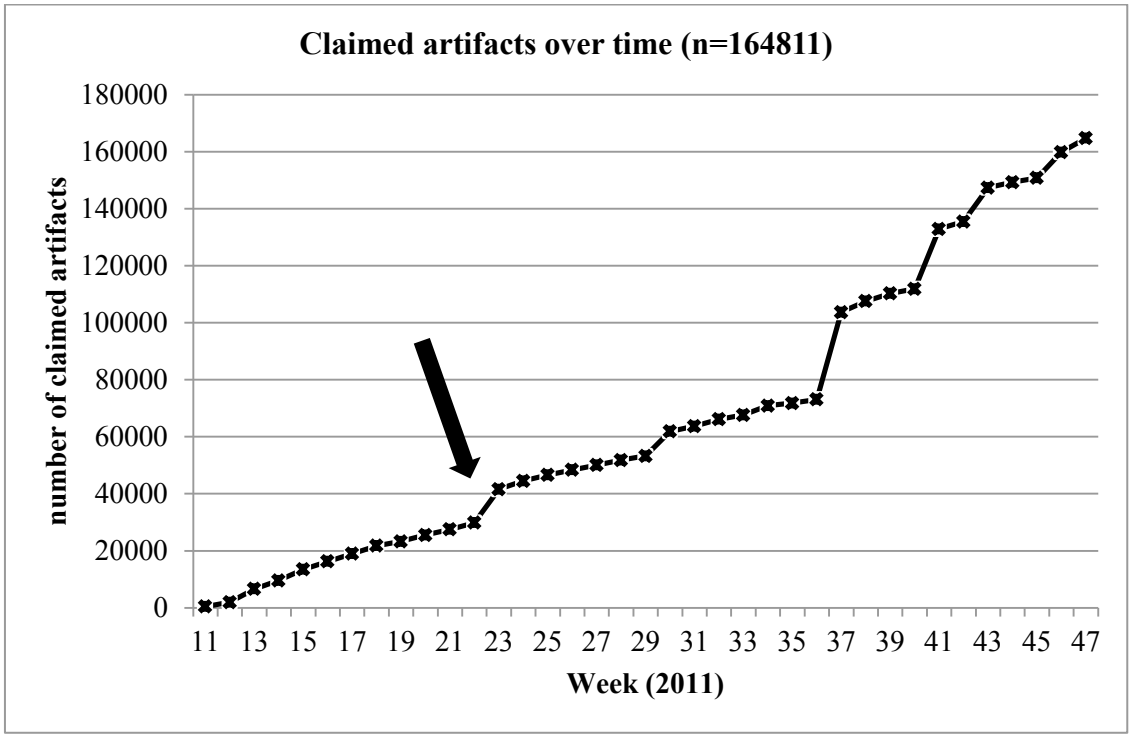


Figure 12: Claimed artifacts over time via the interfaces on INSPIRE. An increased influx of claimed artifacts is due to the mail-out and workflow with arXiv. The arrow marks the starting point of the engagement campaign. The accelerated influx of claimed artifacts every several weeks reflects the results of a mail-out.

The timing of the response after the emails that have been sent out is strongly focused on the 24 hours immediately following (Figure 13). The response pattern continues to show a significant increase in responses after this initial 24 hours.

Studying the researchers who contributed over time (Figure 14) reveals an interesting development: during the weeks in which a mail-out takes place there is not only an increase in the invited contributions, but also there is an increase in researchers who serendipitously discover the tool and claim their publications.

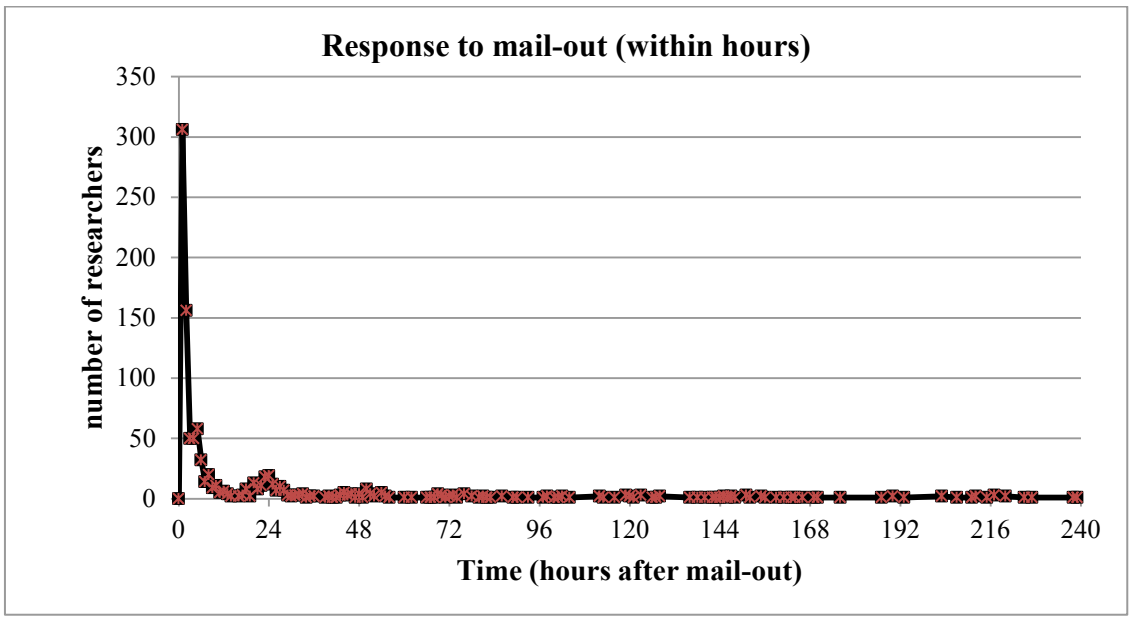


Figure 13: Response to mail-out in hours after the invitations have been sent out (n=1021; the hourly timing was only recorded for the mailings in the end of this test). The majority of the researchers replied immediately after the mail-out, mainly within one hour, after 24 or 48 hours.

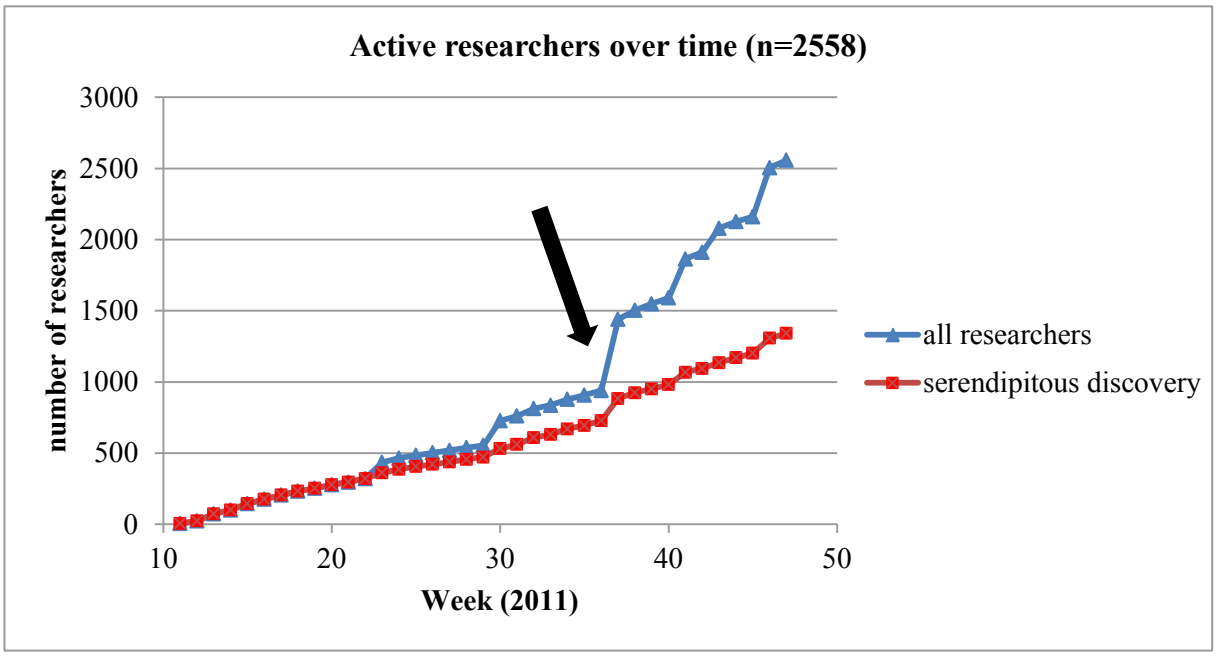


Figure 14: Number of researchers who contributed to the author disambiguation tool over time (cumulative). Two graphs show the overall participation (blue) and the input done after serendipitous discovery (red). The arrow points to an increase in serendipitous discovery during a mail-out. It is speculated that researchers recommend the tool to colleagues.

In addition, a test was conducted to understand the preference of researchers in regard to the workflow they choose: a test of 200 researchers[171] revealed that 82% used their arXiv accounts to proceed (and only 18% proceed as guests).

As the invitation email states, researchers should contact the INSPIRE team directly in case of questions or problems. Indeed, for each mail-out up to 2.5% of the respondents sent full text replies. The majority simply confirmed their publication list or asked for a simple amendment of the list; they often noted that they find the service very helpful. A number of authors encountered problems when using the arXiv login and communicated such technical problems via email. Others referred to more complex issues, for example when names and authorships got mixed up in different clusters. In order to solve such problems, the back office of INSPIRE consulted the researchers and either guided the researchers through the workflow or solved it in the back end. Several times researchers also provided screenshots to illustrate the problem or to suggest improvements for the interface.

This is important to note: they are willing to spend extra time to report problems in order to use the service properly afterwards. The high participation of researchers, their recommendation and inquiries show that researchers indeed perceived the offered tool and workflow as a useful service.

5.4.4 Discussion and impact of the first case study

The response rate of 40% shows that researchers can be engaged in a new tool and service in digital scholarly communication. The engagement was based on the reputation driver found in the research chapter 2 and 3. Moreover, it underlines that in HEP this driver is specifically focused on the publication and citation record and the researchers' visibility within the community and possibly beyond. While this might not be surprising, it is new to use this knowledge in an engagement strategy for digital scholarly communication tools.

The results are further supported by the increase of serendipitous discovery during the week of a mail-out. It suggests that invited researchers spread the word and recommended the tool to other colleagues, students etc. – a process that is targeted by viral marketing in social media. This

[171] All of them having an arXiv account, according to the HEPNames database.

process needs to be highlighted as it shows that the right drivers were placed for participation[172]. The researchers perceived the services as being useful and recommended others to spend time on them. In that respect, the high participation suggest that such tailored engagement strategies should be studied further for other tools and disciplines.

The participation of the researchers who were invited can be considered outstandingly high in comparison to commercial email campaigns (Direct Marketing Association[173]). Comparable numbers in the HEP community are only available from two surveys and thus can only be used to a limited extend for a comparison. For the Parse.Insight survey a response rate of 10% was considered successful (Igo-Kemenes et al., 2010). The survey by Gentil-Beccot et al. (2009) resulted in an estimated response rate of 5% to 10% which was considered "an incredible rate of participation".

The result indicated that several aspects have contributed to the high participation in the new digital scholarly communication service :

- Due to the trusted source, SPIRES and INSPIRE;
    - In several feedback emails users compared both systems, which highlights the trust in the new services and tool and email sender
- Due to the trusted workflow with arXiv that was also communicated in the email and interface;
    - Several feedback emails comment on this workflow
- Due to the drivers communicated in the email;
    - Researchers understood the relevance of the service. This was also evident from the written feedback that has been received. The recommendation of this tool to colleagues underlines this as well.
- Due to an interface that reflects the needs and also translates the drivers and barriers established;
    - Only a few researchers reported that they had problems with the interfaces

[172] The service used for this case study is accessible via the author pages on INSPIRE. The author profile pages have a high visibility with INSPIRE being the central tool for researchers in HEP. Thus, based on this integration, the communication and engagement strategy was associated with a high visibility, as it has been expressed in feedback emails.

[173] http://newdma.org/2012responseratereport [accessed July 31, 2012].

- The correct clustering by the algorithm which has been presented to the users, so that no difficult and complex interactions needed to be done via the interface (see also Weiler, 2012).

Thus, it is evident that many factors contributed to the success of the tailored engagement. The detection and communication of drivers and barriers in building, developing and running this service is present in most of the arguments outlined above. The results are in particular convincing when considering the difficult scenario in which the driver has been studied: a relatively new platform (with the transition from SPIRES to INSPIRE), new service, new interface, and a new email address (INSPIRE) that was used.

In respect of other crowdsourcing endeavors in the library and information science the results cannot be compared. The examples given in Holley (2010) highlight very different approaches and characteristics (e.g. in the user community) in the crowdsourcing initiatives. This also highlights the fact that such initiatives and new (digital library) services, which have not yet taken off in scholarly communication, could learn from this approach that follows a detailed requirements engineering combined with an engagement strategy.

On a final note, it is interesting to refer to a similar, cross-disciplinary service for author disambiguation: Generally, only little data is available that is suitable for a comparison of such services. The AuthorClaim[174] service, running since 2008 for example, reports 100 completed profiles[175]. More details are available on an author identification system by Thomson Reuters: Researcher ID[176]. Searches for HEP researchers in this database reveal that it did not take off in the HEP community[177]. It needs to be speculated why: it appears that Researcher ID had not yet focused on communicating a driver and placing appropriate hooks in its communications with the researchers. It can be hypothesized that it could have been more successful. HEP researchers certainly are willing to participate in such tools, as this study shows. They care about their reputation and are willing to tag (Gentil-Beccot et al., 2009), so they should have had an interest in Researcher ID generally.

[174] http://www.authorclaim.org [accessed August 3, 2012].

[175] http://www.slideshare.net/repofringe/1100-krichel-edinburgh-2011-0803 [accessed September 22, 2012].

[176] http://www.researcherid.com/ [accessed August 3, 2012].

[177] As of July 2012 only a few HEP researchers had signed up for it; some of them also do not appear to be up-to-date or "real researchers" profiles. The search for the keyword „physics", for example, reveals only 161 results [accessed August 1, 2012].

5.5 Second case study: The "hesitation" barrier in the HEP community

5.5.1 Introduction

This chapter describes the second part of the case study in HEP. It focuses on the hesitation barrier which will be studied within the framework of research data sharing.

In this instrumental case study the concept of an embedded research information manager (Walshe, 2011) is used to connect and engage with the HEP community. This approach involves establishing the first contact, and intensifying the collaboration via working groups or triggering of feedback loops, for example. This approach is also based on the results presented in chapter 3, which suggested a strong collaboration with the community. The communication and engagement process focuses on the hesitation barrier in research data sharing and its link to the incentive system. The author of this thesis has been working with the community since the beginning of the PhD (e.g. through the digital library INSPIRE), this concept shall also establish a community awareness that can be sustained in the mid- and long-term.

In the following chapter the chosen methodology and chosen tracks for the enhanced collaboration with the HEP community are explained. Then, as part of this case study, a brief but comprehensive analysis of research data management in HEP is done. This is needed in order to have a common understanding for the embedding and collaboration process. After this the process and results of this case study will be presented. This happens via a detailed analysis of its progress and outcomes. Finally, the discussion will consider whether the approach leads to advancement in the researchers' research data sharing perceptions and habits or not, i.e. if the hesitation barrier can be overcome.

5.5.2 Approach

The concept of embedded librarianship comprises a strong collaboration of the research library or information management service with the customer, i.e. the community. This could, for example, mean that library staff members are placed within the target groups (e.g. the faculty they work with; see Carlson, 2011). In this way the research groups gain enhanced access to information resources and the embedded librarian acts as a "bridge" to specialized information resources. Carlson (2011) highlights that the concept of an embedded librarian is interesting for "information resources as they are generated over the course of the research, such as data […]".

The approach here follows the findings and suggestions by Walshe (2011) who define the role of an "embedded research information manager". They recognize a "gap in support and guidance for researchers in managing data across the entire lifecycle" and recommend that "[…] support services need to engage with all levels of researchers, in order to understand and capture complex needs and align services accordingly". This means that the embedded information manager[178] closely interacts with the community; this could also mean that requirements for information management are gathered and translated into new services, offered by the institution.

In practice this means that in this case study the author of this thesis meets researchers, listens, and transfers knowledge and best practices from other disciplines, so that lessons learnt elsewhere can be reused in HEP. This is particularly important in terms of the known hesitation barrier and significance of the reputation driver. The author of this thesis interacts with the community by considering both themes within the work together. This means that the approach takes advantage of the information gathered in chapter 3 and in particular of the author's involvement in the digital library INSPIRE and her placement at CERN.

It is hypothesized that the close collaboration with the community can help to engage researchers in digital scholarly communication (this will be discussed in chapter 5.8). If successful, the interaction of information management and community could be incorporated into enhanced feedback loops and thus possibly merge into the requirement engineering of tailored services. Depending on the capability of the available information systems (in this case INSPIRE), this then leads to the design and eventually offers of corresponding services. Finally and ideally, this could lead to a "win-win" situation, in which information management knows what services are needed; while the community has support and corresponding services at hand to pursue data sharing. Such results would underline that an enhanced data management support for research communities by information management is needed (cf. Walshe, 2011).

The starting point of the collaboration coincides with the start of this PhD project (December 2009). But the collaboration was intensified over the summer of 2011 and winter 2011/2012. This change was intended and related to the availability of results from previous research (when at least preliminary results from chapter 2 and 3 existed).

[178] In this thesis the terms "embedded research information manager" and "embedded information manager" are used synonymously. The author of this thesis is part of the Scientific Information Service (in this text often called "information management") team at CERN and is involved in the digital library INSPIRE. She conducts this case study.

The strong collaboration focused on the following groups: Study Group Data Preservation and Long Term Analysis in High Energy Physics (DPHEP[179]) and the data preservation task force of the Compact Muon Solenoid (CMS[180]) experiment. The DPHEP study group focuses on data preservation covering all the research data existing in HEP. Its members are researchers and data managers from the HEP community. It is endorsed by the International Committee for Future Accelerators (ICFA) and was founded before the PhD project started (first meeting took place in Hamburg, January 2009). The CMS experiment is one of the four big experiments on the Large Hadron Collider (LHC[181]) at CERN. The collaboration comprises over 3,000 researchers. The data preservation task force was endorsed by the experiment's collaboration board in order to address an urgent need that arose from the ongoing demands for data preservation and access by funding bodies and others.

An additional smaller embedding, closely related to the working groups above, was undertaken in this case study. The "Harmonize data preservation group among LHC experiments group" is an independent informal group and the collaboration is still ongoing (July 2012)[182]. It comprises representatives from all the experiments on the LHC. The group started to meet at the end of 2011, resulting in regular meetings. Here, only a brief overview of the existing progress and results will be given.

5.5.3 Development of a common terminology for research data in HEP

Over the course of being embedded as an information manager in the community, the need for a common terminology became evident immediately. The only common terminology available for an exchange beyond one experiment that existed by the time of the beginning of this case study were the four consecutive tiers described by DPHEP (2009) and briefly described on page 88.

For the embedding process in the community, the following overview chart was generated to summarize the characteristics of the DPHEP levels. It is needed to showcase the complex range of research data that exists in this discipline. Research data in HEP is often referred to as "big data"

[179] http://www.dphep.org [accessed September 8, 2012].

[180] http://cms.web.cern.ch/ [accessed July 8, 2012].

[181] http://lhc.web.cern.ch/lhc/[accessed September 8, 2012].

[182] The CMS data preservation task force demanded an extended discussion of data preservation and data access with the other LHC experiments. The CMS task force commissioned the information management team with the organization of this extended group, called "Harmonize data preservation group among LHC experiments group". Information management took over the role of moderation and knowledge transfer into the group (i.e. from other disciplines or DPHEP).

(Brumfiel, 2011). This of course holds true, but only comprises one part of the spectrum of research data in HEP (Figure 15, left side). It needs to be highlighted that there is a continuum of research data ranging from the "raw experimental data" of high complexity and size to more processed data with a higher level of abstraction and to smaller datasets that are highly processed and, for example, used as supplement to publications. The first can be data in the range of Petabytes size as they are produced by the LHC experiments; the latter are often "only" Kilobytes sized datasets.

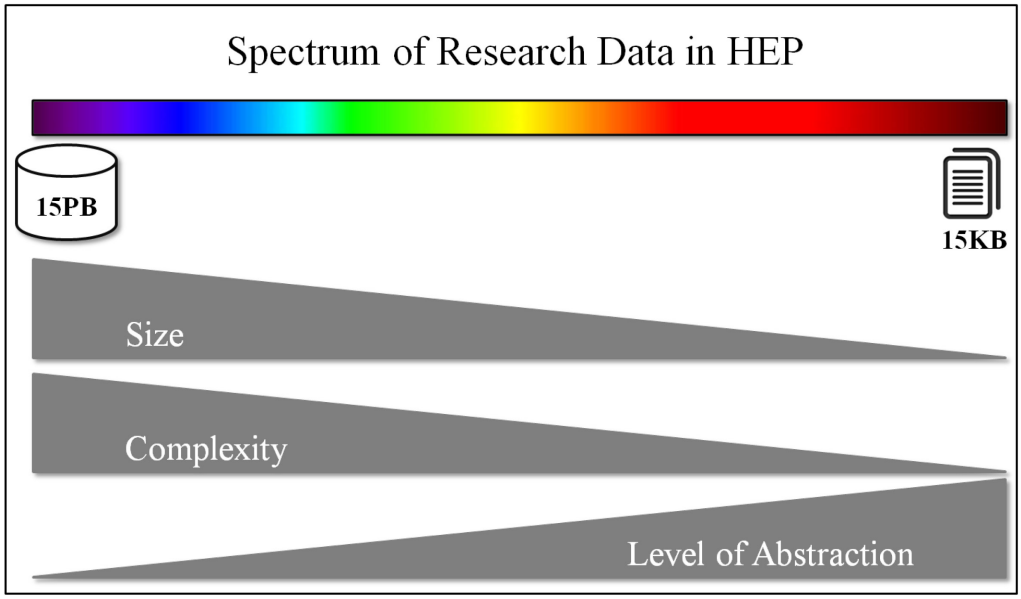


Figure 15: The spectrum of research data in High Energy Physics. From left to right, the size and complexity of the datasets decrease which is associated with an increase in the level of abstraction. It is used as a common starting point for the data-related discussions with the community. This abstract concept has been developed based on the four DPHEP levels described below.

This range of materials makes the field of data preservation and sharing in particular interesting for information management, and highlighted the need for a strong collaboration with the community to solve the issue of what to preserve, where and how. This is not trivial, given the complexity and size of some datasets, and the existence of experiment-specific virtual research environments and access demands. The latter also points to the challenge of data sharing, where it is needed to investigate how data can be shared so that it is reusable for others.

The four tiers of research data as defined by DPHEP (2009) are given in the following. Subsequent levels are inclusive:

- Level 1: Publications-related information search. The preservation model foresees the provision of additional documentation, such as additional data tables or other high-level products, e.g. notes, codes. In the spectrum (Figure 15) this level can be found on the right.
- Level 2: The preservation model foresees the preservation of the data in a simplified format for outreach and simple training analyses.
- Level 3: In order to provide full scientific analysis potential based on existing reconstruction, the preservation model foresees the preservation of the analysis software and data format.
- Level 4: The most complex layer aims at preserving the full potential of the experimental data. Thus, reconstruction and simulation require the software and base-level data. In the spectrum (Figure 15) this level would be on the left.

This concept was the starting point for the discussions in the individual collaboration tracks with the HEP community. The common terminology turned out to be of significant value when starting any discussion about the data in any working group or task force, as the nomenclature directly pointed to the position on the spectrum, and the respective challenges of the particular type of data became clear.

5.5.4 Results from the second case study

At the beginning of this case study it became evident that only limited action had been undertaken in regard to data preservation and data sharing. If they existed, solutions were not widespread, and focused on a single experiment or a single case study, and did not happen on a general layer beyond one experiment. This is also evident in the interview with Peter Igo-Kemenes (Dallmeier-Tiessen et al, 2011d) conducted in spring 2011, in which he discusses the challenge to recover and preserve data from old experiments. He highlights that there is an emerging awareness indicated by the advent of the DPHEP initiative.

Since the beginning of this case study, the dynamic situation in research data sharing has changed across disciplines. The topic of preservation, sharing and access to data has increasingly been discussed, and has also raised awareness and demands at the highest European political levels (see for example European Commission, 2012a; Kroes, 2010). This overall development has to be noted to understand the general influence and pressure that accompanied and also influenced the development in HEP. The influence has also been seen in the discussions within the individual

embedding tracks. Frequently participants in the task forces reported conversations with funding agencies or their institutions, requiring actions to promote data preservation and sharing.

Participation in the individual working groups or task forces went through several different phases, developments and milestones. These events during the collaboration with the community were mainly triggered by a demand for wider exchange with the community beyond the given task force.

In the following the results are presented. This overview distinguishes what kind of developments have been observed, supported and triggered. This is done for the individual "collaboration tracks" that have been followed, namely the DPHEP group and the CMS task force. Finally, services for data preservation and sharing on INSPIRE are presented. These have been developed based on the new enhanced feedback cycles that emerged through the collaboration.

5.5.4.1 Study Group for Data Preservation and Long-Term Analysis in High-Energy Physics

A first contact with the DPHEP community was established with the very beginning of this PhD project at CERN when the third DPHEP workshop took place at CERN. Over the course of the embedding a growing awareness within the DPHEP community in regard to data preservation and data sharing could be observed. In particular the topics associated with the exchange and reuse of research data gained relevance, in particular OA to data. Researchers were rather hesitant in that regard in the beginning – according to their statements due to the complexity and size of the data output in HEP. But, frequently participants in meetings reported from requests from funders and emerging policies, requiring actions to advance data preservation and sharing.

The information management team at CERN contributed a presentation to the DPHEP workshop in 2011[183] at Fermilab (US). Best practices and experiences in data preservation and sharing from other disciplines were presented. This knowledge transfer highlighted, for example, studies that investigated citations increases with data sharing, a statement that directly points to the "incentive" system and reputation. Also, examples and best practices from other disciplines were shown. It was pointed out that there is a need for action in HEP as the community has been slow to act in comparison to many other disciplines.

[183] https://indico.cern.ch/conferenceDisplay.py?confId=116485 [accessed September 8, 2012].

The latest milestone was passed in May 2012, when a DPHEP meeting[184] in New York took place, again with a presentation by the information management team, which highlighted best practices from other disciplines. That presentation also showcased new data sharing and preservation features that would be offered via the digital library INSPIRE.

At the meeting in New York in 2012 the "Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics" (Akopov et al., 2012) was published, comprising a comprehensive analysis of the HEP data preservation landscape. The authors of this detailed report are mainly physicists or IT-related staff members, and representatives from the information management.

The influence of the continuous knowledge transfer (by the information management team) can be seen: the report highlights that "in the last year an encouraging tendency to initiate concrete projects within the participating experiments and laboratories has been observed" (Akopov et al., 2012). This points to intra-HEP collaborations, e.g. with INSPIRE, and also to initiatives beyond the HEP discipline. Furthermore, it outlines emerging initiatives on different levels, from the experimental to the international.

The strong collaboration triggered an extended view beyond the HEP discipline. This can be traced back to the intensive knowledge transfer and reporting on best practices from other disciplines. The authors note that the report is seen "as a first step in a new period where data preservation in HEP will develop at international level, with strong synergies with other scientific fields and with the ambitious goal of enhancing the potential of the HEP data by explicitly using a global, long-perspective and flexible access approach" (Akopov et al., 2012).

The enhanced collaboration resulted also in closer feedback loops for the development of new tailored services for the community. The resulting services are considered necessary for future work: the digital library INSPIRE is mentioned as a practical solution to provide services in data preservation and additional scholarly materials within HEP (see quote below). The assignment of Digital Object Identifiers (DOI) is highlighted with a suggestion that INSPIRE should provide this service as well[185]. These services point to the drivers and barriers identified, in particular in regard to the incentive system.

[184] https://indico.cern.ch/conferenceDisplay.py?confId=171962 [accessed September 8, 2012].

[185] See footnote on page 13 of Akopov et al. (2012).

> *"Technologies for data preservation are investigated such as virtualisation and virtual repositories, data and analysis migration procedures, data validation suites and archival infrastructures. The management of information and its storage is also examined, including the extension of documentation in the public domain and the enhancement of information by storing figures, data, notes and internal legacy material in collaboration with the INSPIRE service."*
>
> *Quote from the executive Summary of the DPHEP report (Akopov et al., 2012)*

This shows that by collaborating with DPHEP from early onwards, the services were placed prominently within the initiative and, more importantly, workflows were defined together that will offer enhanced preservation services. According to the DPHEP initiative these services will be developed further in the oncoming years together (Akopov et al., 2012).

5.5.4.2 CMS data preservation task force

The collaboration with this task force comprised the participation in regular meetings, and the preparation of information materials for the group and for the policy that was being developed by the task force. In particular, knowledge transfer from other disciplines to HEP was sought in this task force. The group started its meetings in February 2011. The two participants from information management joined the group in summer 2011.

The milestones of this collaboration are mainly related to the internal discussion and approval processes in the experiment. The hierarchical structure in the experiment requires several committees to approve an initiative that affects research practices within the experiment. This means that the task force started working on a draft policy for the experiment that went through several feedback loops within the experiment before being approved. It resulted in a CMS-wide policy, namely the "CMS data preservation, re-use and open access policy" (CMS collaboration, 2012).

The development of the policy and associated documents was heavily influenced by the political framework, i.e. funders. The increasing number of policies by European funding bodies pointed to specific topics (e.g. OA to data) that were then emphasized in the preparation. They were also highlighted by the task force leader in her presentations to the CMS collaboration.

Overall, hesitation among the researchers beyond the task force was evident, when presenting and discussing open sharing of data. This is related to the strong internal collaboration and effort that is needed to produce and analyze data. Thus, data is a scientific output of highest value and researchers hesitate to share with others who have not contributed to the experiment. A strong fear of misuse appeared prominent as well as concerns in regard to liability. A demand for reuse tracking was made when sharing data. The aspect funding frequently emerged in the task force, mainly associated with the need to cover the emerging effort for data preservation, sharing and corresponding services.

During the course of the task force, the information management team mainly contributed to the open questions on the experiments side in terms of access models, embargoes, licensing, etc. The latter particularly addressed the observed fear of misuse and resulting liability questions. This support comprised knowledge transfer from other disciplines, such as molecular biology or earth sciences, which have similar policies already in place. Also support in regard to the funders' expectations was given. In addition, the information management team contributed by offering and designing practical solutions to the individual policy steps (in particular for data preservation on levels 1 and 2). This in particular pointed to the demand for reuse tracking that could be facilitated via the digital library INSPIRE (and the assignment of persistent identifiers to research data therein).

In conclusion, the information management team at CERN developed some of the specifications in the discussions and in the resulting policy document. The policy explains that the content has been developed in the light of the increasing demand for public access by "funding agencies" (citation below) and focuses on the possibility of reuse of CMS data.

> *"CMS upholds the principle that open access to the data will, in the long term, allow the maximum realization of their scientific potential. To that extent, CMS will provide open access to its data after a suitable but relatively short embargo period, allowing CMS collaborators to fully exploit their scientific potential.*
>
> *This policy describes the CMS principles of data preservation, re-use and open access, as well as the relevant actors in all these tasks and their roles and responsibilities. CMS understands that in order to fully exploit all these re-use opportunities, immediate and continued resources are needed. The level of support that CMS will be able to provide to external users depends on the available funding. This policy addresses the moral*

> *responsibility of CMS for its data, as well as the increasing concern of funding agencies worldwide and the civil society for the preservation and re-use of scientific data."*
>
> *Quote from the introduction of the CMS policy on data preservation and access. (CMS collaboration, 2012).*

The policy describes preservation and access as two separate but linked topics. It introduces an embargo period for public access to some data levels, so that the experiment is able to exploit the dataset before the public release.

In more detail, the policy describes preservation and access to the individual data levels (according to the DPHEP levels, DPHEP 2009). This is the first time that practices for data sharing have been formally agreed on within a HEP collaboration. This is especially remarkable as Level 3 data (citation below) is considered as detailed data that allows reconstruction and reanalysis.

> *„Level 3: Reconstructed data and simulations, together with the software, analysis workflows and documentation needed to access the data, understand them, reproduce published analyses, perform new analyses not requiring re-reconstruction of the data or new simulations.*
>
> *CMS level 3 policy: CMS preserves the reconstructed data by forward-porting i.e. by keeping a copy of the data reconstructed with the best available knowledge of the detector performance and conditions available. These data include simulation and can be analysed with the central CMS analysis software but cannot be re-reconstructed. Analysis procedures, workflows and code are preserved as part of the CMS code repository under the responsibility of the CMS Offline project. Responsibility for archiving of the data sits with the present tiered structure of the CMS computing infrastructure."*
>
> *Quote from the detailed access plan to Level 3 data in the CMS policy on data preservation and data access (CMS collaboration, 2012).*

The policy by the CMS collaboration underlines that Level 1 to Level 3 data is shared OA. Level 4 data is not considered feasible for sharing (CMS collaboration, 2012). The impact of the strong collaboration of information management and researchers can be seen in the policy which hints towards possible workflows and facilitators for the individual preservation and access levels. For Level 1 and 2 the information platform INSPIRE will facilitate the corresponding services,

including reuse tracking. The clauses in regard to the liability and reputation of the collaboration are incorporated using standard licenses, such as the waiver CC0[186] as they had been reported previously from best practices in other disciplines (Hrynaszkiewicz & Cockerill, 2012).

In summary, the results from the enhanced collaboration with this task force show that the information management service supported and facilitated important developments within the community. The information management team used knowledge transfer via best practices from other disciplines. The input benefited also from a network of funding bodies, other libraries and data centers, so that topics and best practices in terms of policies, access models, licensing, etc. could be addressed. Corresponding services for the digital library were offered and developed[187]. It needs to be highlighted that the strong collaboration resulted in the removal of the hesitation barrier. It is based on a continuous interaction and cross-fertilization. This means, both sides benefited from the collaboration: tailored services for the community that link to the task force's demands and the communities' incentives were developed. Linked to this, a long-term collaboration was established including enhanced feedback loops for further development in the digital library.

The collaboration with the „Harmonize data preservation group among LHC experiments group" is still ongoing, but overall the preliminary results are very similar to the ones described above. The information management team influenced and triggered discussions among the representatives that were then carried back to the individual experiment where further internal discussions took place.

The group started with this discussion of the CMS principles and extended their discussions to the individual challenges the experiments face. One major milestone was reached in May 2012, when the individual experiments reported their progress publicly during the "International Conference on Computing in High Energy Physics" (CHEP). The presentations[188] highlighted the existence of a common approach. One can speculate that the resemblance of the structure and DPHEP level approach in regard to preservation and access is due to the regular "harmonize" meetings. The progress shows that this collaborative approach in framing awareness worked successfully.

[186] Public Domain Dedication, http://creativecommons.org/publicdomain/zero/1.0/ [accessed July 8, 2012].

[187] Through the extended discussion of the policy with the other collaborations in the "Harmonize data preservation group among LHC experiments group" and the other embeddings (and keeping in mind the barrier "standards/interoperability") it has been ensured that standard and interoperable services are developed, but that such tailored services serve the wider community.

[188] http://indico.cern.ch/conferenceDisplay.py?confId=171962 [accessed August 3, 2012].

5.5.5 Discussion and impact of the second case study

The purpose of this case study was to study the hesitation barrier and its characteristics in HEP. The hesitation barrier in HEP is reflected, for example, in the prominent fear of misuse when discussing potential access options for the policy papers. It is evident when considering the relatively little experience with open data sharing so far. Given the enormous data output of the community, only few open sharing services are available. But the progress and outcomes of the embedding have shown that this barrier can be surmounted. This happened, for example, by offering solutions that connect to one of the known drivers in HEP: citations. The strong collaboration was able to convey relevant services, such as DOI registration and data citation that can help building trusted workflows, and the framework for an incentive system which supports data sharing. The policies and recommendations in general, but in particular the notion of particular services (for example DOI registration), highlight the fact that the collaboration can be considered fruitful in the mid- and long-term.

Interestingly, the theme funding (as it has been detected in chapter 2 and 3) also emerged frequently during the embedding. Researchers noted that upfront, mid- and long-term investment is needed to undertake some of the actions mentioned in the policy (CMS) and report (DPHEP).

Again, it has to be noted that the surrounding framework also changed during this core period of time (about a year). The influence of the increasing demand for data access by funding bodies cannot be estimated highly enough. The role of information management in the collaboration also comprised a continuous update on such requirements. This is reflected in the specific output, such as the policy, takes into account the developments and the specific demands: management plans, preservation, access and licensing etc. With the embedding, such an immediate and direct information support was given that – together with the transfer of best practices from other disciplines – helped to establish the awareness.

It is also necessary to reflect on the decision-making process of an individual researcher as against the community or a group which makes a decision. This is particularly evident in HEP, where experimental data is often produced by big collaborations like CMS and where a complex serious of interrelated governance bodies make the decision to share or not. The embedding has shown that this does not change the driving factor, reputation, and the barrier, hesitation.

The results show that the strong collaboration with the community enhanced the understanding of the partners on both sides. The community trusts the information management in designing and

offering corresponding services, and the information management team received feedback on what kind of services are needed, proven by the policy and recommendations referenced above. The one-to-one situations allowed the establishment of a sustainable collaboration that will help designing the service portfolio in the future. The work with the community and the design of the services thus also enables individual community practices, by building on best practices from other communities.

The embedding helped to enhance the development cycle of services for the digital library INSPIRE. The information management team proposed services based on the conversations within the task forces. Based on their initial feedback, such services are developed and deployed along with the sharing of datasets. The information management team demonstrates them to the community to kick off further sharing and feedback.

This short feedback cycle leads to the design and implementation of tailored workflows for the information platform INSPIRE. They are seen in the development of the following services[189]:

- Individual Machine Readable Cataloguing (MARC[190]) records for research data
- DOI registration for research data and possibly other scholarly materials (Figure 16)
- Reuse tracking, facilitated via the above (Figure 17)
- Authorship of research data
- Display of research data citations on the author pages on INSPIRE

These workflows and services follow the community's demand for an incentive system for data preservation and data sharing. Moreover, they are part of the policies and recommendations already mentioned, which affect the majority of the community. It thus could be assumed that these services will indeed be used by the community. The services feed into the global framework of DataCite which facilitates data citation across disciplines[191].

[189] That are available via the digital library INSPIRE.

[190] http://www.loc.gov/marc/umb/ [accessed August 8, 2012].

[191] Thus, it implicitly addresses the theme of standards and interoperability as well (chapter 3).

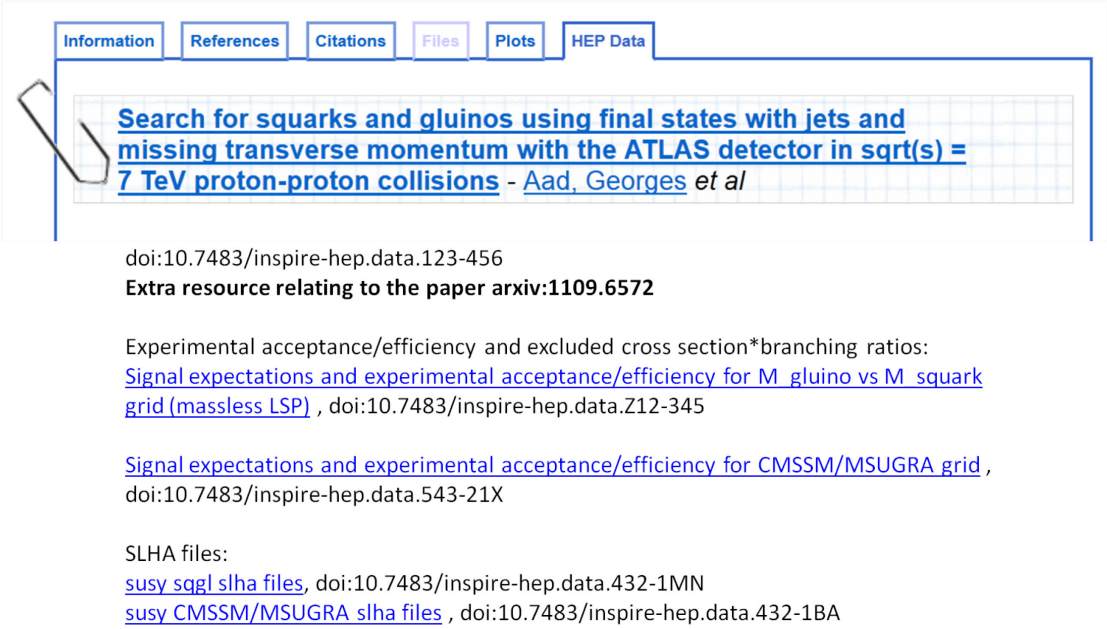


Figure 16: The digital library INSPIRE showing data citation via DOIs (Digital object identifier). This is the prototype design that is being used to sharpen the design with the community.

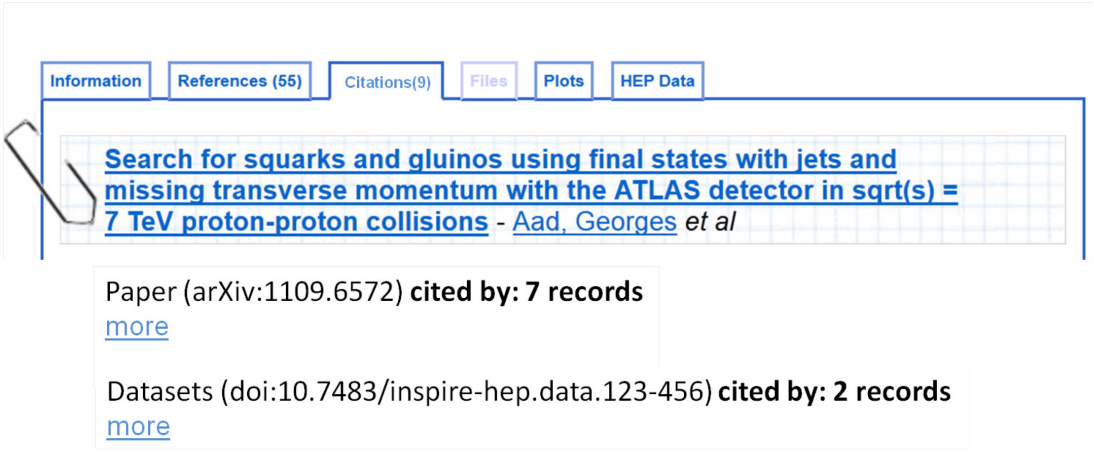


Figure 17: The digital library INSPIRE showing data reuse tracking. This is the prototype design that is being used to sharpen the design with the community.

In conclusion, the hesitation barrier was prominent in the HEP community. However, embedded community engagement helped to take the first steps in surmounting it. The result is based on:

- The engagement with the community: in particular the opportunity of informal feedback loops on policy development, services development, INSPIRE development
- Communication of best practices and lessons learnt from other disciplines
- Offering trusted workflows that connect with current incentive mechanisms (data citation, a mechanism that is now incorporated in most of the HEP policies described above)
- Collaboration with other stakeholders, e.g. funders, libraries, data centres and thus immediate input into the policy building processes

5.6 Summary: Review of drivers and barriers in these case studies

The instrumental case study within the HEP community comprised two independent parts. Both study the reputation driver and the hesitation barrier which are part of the societal layer. They showed that they are of high relevance and can be used in engaging researchers in new tools and services. The second case study focusing on research data sharing refined the hesitation barrier on a disciplinary level, and showed that information management can work successfully with the community to surmount it. In the framework of digital scholarly communication, this means that HEP researchers are willing to participate in new tools, services or workflows if the right drivers and incentives are implemented and communicated effectively.

In this case study many of the themes that were detected in chapter 2 and 3 appeared in one way or the other as well. For example, the impact of legal and policy frameworks (i.e. funder requirements) and also standards & interoperability (i.e. DOI as a standard) emerged during the interaction with the community. The theme funding emerged frequently as well, in particular in the embedding. These themes have not been addressed in this case study and need further investigation.

It is interesting to consider the themes reputation and hesitation within HEP in the wider framework of the results from chapters 2 and 3. They underlined that both themes are strongly connected to the current incentive system. The case study proved that this impacts the daily practices and habits of researchers. In HEP, publications and citations are highly significant for researchers. Both contribute to the researcher's reputation, which is relevant for new jobs or grants. This is in parts driven by current academic incentive and research assessment schemes, e.g. in the UK with the Research Excellence Frameworks demanding four publications from a department

member for the assessment[192]. In Germany, the DFG asks for the top five publications of researchers when submitting their proposals[193]. In consequence there is a focus on the output that is counted, in proposals, assessments or selection committees. This is the driver for engagement and in consequence a barrier for new emerging workflows in Open Science if not incentivized correctly.

But there are emerging trends that have begun to appear at the time of writing, e.g. European Commission (2012a) announced that one should now study ways in which data sharing could be incentivized and integrated into funding schemes so that Open Science will be supported.

In more detail the European Commission recommends to its member states in July 2012 the following:

> *"...adjusting the recruitment and career evaluation system for researchers and the evaluation system for awarding research grants to researchers so that those who participate in the culture of sharing results of their research are rewarded."*
>
> *"Improved systems should take into account research results made available through open access and develop, encourage and use new, alternative models of career assessment, metrics and indicators."*
>
> *"...institutions responsible for managing public research funding and academic institutions that are publicly funded assist in implementing national policy by putting in place mechanisms enabling and rewarding the sharing of research data."*
>
> *Quote taken from European Commission (2012a)*

These citations illustrate the emerging awareness concerning the changes needed in the incentive system so that the full potential of digital scholarly communication can be used. These case studies have underlined that there is a need for a link to the incentive system to pave the way for Open Science.

[192] See also chapter 4 for details, p. 62.

[193] http://www.dfg.de/service/presse/pressemitteilungen/2010/pressemitteilung_nr_07/index.html [accessed August 12, 2012].

5.7 Applicability of case studies to other disciplines

The HEP community and its scholarly communication landscape is a specific case. This is mainly due to its relatively small size with centralized (large scale) experimental and scholarly communication facilities.

There are advantages in the HEP community, for example, the overall willingness to participate in tools, as already stated by the researchers in 2007 (Gentil-Beccot et al., 2009). It has also to be noted that the HEP community offers the advantage of a "one stop shop" digital library, which might not exist in other disciplines, and the absence of which might make an engagement strategy more complex in other domains. This means that the case study can be considered limited as many disciplines do not have such centralized information systems. Many information providers (i.e. repository providers) do not have easy access to their respective communities, which hampers an engagement strategy.

The approach and results might thus look different in other disciplines, from the kind of platform that is being used to the engagement strategy and the processing of the workflow and results. But overall the results need to be considered as an example of how the identification of drivers and barriers can help to work with the communication and development of new services in scholarly communication. The cross-fertilization can happen on different scales, from small scale endeavors to large-scale projects. An increased effort for such collaborations might be compensated by the engagement of the community, for example in crowdsourced tools.

It needs to be pointed out that one uses the reputation driver due to its significance. It has been shown that the driver exists and triggers the expected participation. However, it is crucial to keep this significance in mind when developing the services, workflow and the engagement strategy. The quality of the developed services needs to be in correspondence to the significance of an individual's reputation. There might only be one chance to engage the researchers - and if the service and tool do not work as expected, additional future engagements might be at risk.

In regard to research data and the hesitation barrier, the experience should be easily applicable to the other communities. Even though the infrastructural framework and the research datasets might look different, researchers in other disciplines also face similar challenges that emerge from the increasing number of policies and requests by funders, e.g. the data management plans required by

the NSF[194].This means that strategic support on how to deal with the demands might well be welcomed.

Despite some limitations, it has been shown that both case studies in HEP use workflows that impact scholarly communication as a whole, and significantly on a global level. The first case study (chapter 5.4) uses a workflow that merges into the global ORCID initiative to study the driver reputation. The second one (chapter 5.5) resulted in the implementation of services that merge with the global DataCite[195] initiative. The studies and engagements have been conducted on a disciplinary layer, but transfer to a cross-disciplinary layer in these global frameworks. In addition, the results build an understanding and framework for content recruitment and crowdsourcing in repositories and digital libraries. The case studies have ensured an enhanced participation of researchers in the digital library and also paved the way for the input of additional materials such as research data. Such content recruitment is needed to acquire a comprehensive scholarly record in the digital library – not only in HEP, but also in other disciplines.

This implies that the results obtained are of high relevance for other (discipline-specific) stakeholders who may be interested in providing analogous workflows. This was a test using the unique opportunities offered by the HEP community as a case study. Of course, when working with other communities one will need to understand the adaptations needed, what facilities and workflows to use, and to combine this analysis with a detailed understanding of the communities' drivers and barriers in digital scholarly communication. From chapters 2 and 3 it became evident that drivers could look very different in a particular discipline – in the humanities for example, in which monographs or books are the main publication outlets and datasets look completely different. The cross-disciplinary analysis (chapters 2 and 3) highlighted the diversity in that regard and can be a starting point for further disciplinary case studies.

But given the overall framework of research assessments and associated incentive systems (see results from chapter 2 and 3, and 5.6), it is expected that the relevance of the reputation driver in the HEP community can also be found in many other disciplines and thus could be used to steer a researcher's interest in particular tools in scholarly communication. Disciplinary characteristics would – of course - apply as well.

[194] http://www.nsf.gov/eng/general/dmp.jsp [accessed August 2, 2012].

[195] http://www.datacite.org [accessed August 12, 2012].

For the HEP community, it has been shown that "hesitation" hurdles in regard to data sharing can be overcome by a tailored approach for a specific community. The strong collaboration of information management and the community allowed for the design of corresponding services that simplify and incentivize sharing. This endeavor shows that a researcher's reputation and links to the incentive system will be important factors to consider in any actions to be undertaken – beyond HEP.

5.8 Lessons learnt for the role of information management

The role of information management within the two case studies is new, as it involved an active bidirectional communication, collaboration and in particular engagement with the community. This results in tailored services on the information management side (as has been explained with regard to the shortened feedback loops in chapter 5.5.5 and 5.6). Thus, the results of this case study are particularly interesting for information providers or information architects and other stakeholders who are active in scholarly communication. Stakeholders involved in providing platforms, repositories or any related services can build on this case study and use the drivers and barriers (detected for their respective user groups/communities) for the advancement of their information services.

The new role comprises the engagement of researchers within a virtual research environment (first case study with INSPIRE, chapter 5.4), and the physical embedding into the researchers' environment, in dedicated task forces, working groups, one-to-one interactions, etc. (second case study, chapter 5.5.5).

The report by Walshe (2011) on the role of an embedded research information manager pointed to a lack of support and guidance for researchers in managing data. According to the authors, in the recent years among librarians (in an academic setting) there is a move towards "liaison or outreach roles, more pro-activity in engaging with researchers and ultimately toward embedding information support within research teams". This case study indicates that this approach works in practice. These new role models result not only in a refined knowledge about drivers and barriers, but also in surmounting such barriers and enhanced services. In the first case study (chapter 5.4) the response rate of 40% shows that information management managed to get HEP researchers engaged in this service. The results provides evidence that with the support of information management in design, implementation and engagement strategy, researchers are willing to

participate in digital scholarly communication and hesitation can be surmounted. This implies an identification and design of the correct hooks and incentives by information management.

In addition, the results are especially relevant to staff members of digital libraries, as they show that information service providers can crowdsource metadata services from the community, provided that quality controls are implemented[196]. In this case study the workflow is quality assured by editorial staff members, although the bulk of the incoming information can be processed automatically via the workflow with arXiv. So in fact these results can reroute the burden of emerging services and tools from the limited availability of resources in information management to the "crowd" (researcher). Such results are of relevance for the information management services that wish to contribute to global initiatives such as ORCID, which builds up a global identifier system for researchers. Here, it has been shown that workflows can be built that limit the burden on the information management side while providing high-quality data for a global initiative.

The results support the hypothesis stated at the beginning that information management can play a role in surmounting barriers, e.g. establishing a data sharing culture. It has been proven that information management can support a research community, i.e. to raise awareness in data sharing in HEP. It has been shown that this moves beyond a mere information transfer, but can also comprise the design and creation of corresponding services. In this case study the implementation of the data citation workflow and thus the implementation of the link to the incentive system can be considered a key in removing the barrier.

The development of this data citation workflow is of the utmost importance, in particular for future activities in this field. Given the emerging awareness and demand at the highest political levels (see chapter 4 and 5.7) it can be expected that cross-disciplinary approaches and a corresponding incentive system will be seen at national and international levels. If implemented, such new systems will take data sharing and data citation into account, so that the establishment of such a workflow in HEP needs to be seen as work in progress, with follow-up services to come.

In this relatively new domain of data preservation and data management there are numerous roles not yet clearly defined. Even though training exists, more detailed specializations will be required over time (cf. Pampel, Bertelmann, & Hobohm, 2010). But this career path does not only require specialists, but also comprises coordinating and knowledge transfer responsibilities, as has been seen in the case study. Both have been important assets in the case study.

[196] See first case study in chapter 5.3.

It is nevertheless important to highlight that this analysis focuses on a disciplinary case study and a disciplinary information provision and infrastructure. Information architects and managers who work on an institutional level are faced with different challenges. They usually work with a broad spectrum of disciplines that have a wide range of needs and habits. This also applies to their publishing habits as well as their usage of information resources. Further studies are needed to understand how these results apply to such a diverse environment.

6 Summary and Outlook

In this thesis a comprehensive picture of digital scholarly communication today[197] was built in order to understand drivers and barriers therein. In particular, researchers' participation, attitudes and habits have been studied across disciplines. The results frame a new understanding of the challenges existing in digital scholarly communication. Hypotheses and recommendations have been made to address the challenges. They are later studied in more detail in two case studies in the HEP community.

At the beginning of the thesis the framework of research in a digital environment and scholarly communication was laid out. Particular emphasis was given to the opportunity of Open Science. The need for a better understanding of researchers' attitudes and practices, as well as driving and hindering forces became evident. Two major challenges were chosen for the detailed research questions: OA and research data sharing.

The results showed that the potential of Open Science is not yet fully exploited by researchers. A positive attitude or awareness has been recorded in the study, but this is not reflected in the practices.

Quality and prestige is a deciding-factor for choosing a publication outlet. The perceived lack of quality and prestige of OA journals is a strong barrier to OA publishing. The latest developments show that this barrier might lose weight over time as more and more community-approved and prestigious OA journals are established. The survey also highlighted funding as a strong barrier. The discussion showed that mechanisms on disciplinary and national levels are under way to overcome this barrier.

The interviews provided insights into the different attitudes and practices in research data sharing in the individual disciplines. The advancement and the experience with data sharing are very diverse and interviewees highlight the different disciplinary characteristics of data and associated workflows. Researchers hesitate to share data in many disciplines. This is strongly connected to the lack of incentives in current research assessment schemes and the way research funding is structured today.

Both research tracks, on OA and research data sharing, were combined to a comprehensive picture on drivers and barriers in digital scholarly communication (chapter 4). The analysis allowed the

[197] The research has been conducted over a time period of 3 years between 2010 and 2012.

identification of overarching themes, among which a societal layer appeared to be dominant. It describes a strong hesitation in participating actively in Open Science. In addition, monetary and strategic layers, as well as infrastructural and technical layers have been discussed. They are all strongly interconnected. Furthermore, the results state that generalizations across disciplines are difficult due to the different disciplinary workflows and advancements. Nevertheless, strategies for improvements that were discussed show that it is necessary for stakeholders to address the challenges collaboratively. The overarching themes highlight specific challenges for the individual stakeholders, researchers, infrastructure providers, publishers, librarians etc., but also point to the need to expand the individual horizons to other topics, disciplines or countries. The results of this thesis can be summarized with a quote from one interviewee who states that "there is a whole framework that needs to be changed". This statement points to interconnected drivers and barriers and the need to incentivize Open Science.

These findings have been underlined by the results from the case study in the HEP community. Here, two dominant themes from the societal layer - reputation and hesitation - were studied in practice. They showed that it is possible to overcome barriers and engage researchers in sharing and in the participation in Open Science tools and workflows. This is strongly dependent on the involvement of information management services and a strong collaboration with the community. As a service provider, information management listened to the community's requirements, built tailored workflows and implemented them - with continuous feedback through the collaboration. Such positive synergies were used to tailor the workflows for data sharing. These are new aspects of the role for information management: using new feedback and communication cycles with the community to develop tailored services that are considered useful and thus are used.

It has been noted, however, that the framework of the HEP case study might be considered privileged in comparison to other disciplines or institutions as access to high-level groups in the community was granted to the author of the thesis. Nevertheless, the results of this thesis identify similar challenges across disciplines, so that such support mechanisms can equally be a successful experience for other stakeholders on a disciplinary or institutional level.

The case study exemplifies the potential for overcoming barriers. Researchers were successfully engaged in workflows that facilitates the way for Open Science. This potential needs to be found, defined and used by the individual stakeholders, for example by repository providers that envision more submissions to their platform. But the results underline that such strategies also comprise

advice on legal aspects[198], or standards[199]. It also highlights that joint efforts need to be considered for such activities.

Special emphasis in this thesis has been given to discussion and workflows that link to the incentive system. The latest developments[200] have shown that parallel strategies exist to link new digital scholarly communication services with the current incentive system.

In summary, the results in this thesis have shown that scholarly communication is in transition. Taking into account disciplinary differences, overall awareness and attitudes are positive. But the full potential offered by a digital and possibly open environment is not yet being exploited. The results show that there are multifaceted reasons – with a prominent societal layer – that suppress fundamental changes. But the results underline that there is the potential for profound changes in scholarly communication: the awareness is in the researchers and other stakeholders' minds, in regard to OA publishing and also research data sharing. But practices are often not established.

However, in both research tracks some examples of best practices were detected, e.g. in disciplinary layers. Moreover, the case study showed that such barriers are surmountable, in particular with an enhanced engagement strategy and a strong collaboration of community and information management. Nevertheless, it is evident that many barriers need more effort to be overcome, e.g. when considering licensing and legal frameworks. This also applies to the funding framework. The current funding structure focuses on projects with limited durations. It needs to be investigated how dedicated funding streams, for example, can support Open Science infrastructures and services. Funded research projects need to foresee budgets for OA, data preservation and sharing. Accordingly, it is also needed to incorporate Open Science in the corresponding assessment schemes. The results presented here underline that new workflows, for example data sharing, need to find a link in the academic incentive system.

[198] As it has been seen in regard to licensing for the policies in the second case study on data sharing (chapter 5.5).

[199] Advice and implementation has been given on ORCID and DataCite services for example in both case studies (chapter 5.4 and 5.5).

[200] In the framework of OA for example, new initiatives address the quality and prestige challenge in that regard. With the new eLIFE journal [http://www.elifesciences.org/about/, accessed August 17, 2012] one example has been mentioned. Examples like PeerJ [https://peerj.com/, accessed September 26, 2012] address the themes funding and prestige at the same time. In the framework of research data also dedicated data journals that facilitate enhanced sharing have emerged (see also chapter 3 and 4). Documentation and datasets are made citable objects. In this way data sharing is incentivized. For digital libraries and repositories, similar developments are established that connect repository submission to enhanced services and the incentive system. This could comprise for example a publication list on a personal website (on the institution).

Recent movements in policy frameworks suggest that the announced future developments[201] will further the transition in that regard. Work on other barriers in the future is needed to build on more and better collaboration of stakeholders.

Furthermore, the findings in this thesis do point to the need to conduct further research in the field of Open Science. There is only a limited understanding of the dynamics. This applies in particular to research data sharing, but also to the open sharing of other scholarly materials. The concepts and frameworks of their integration into trustworthy virtual research environments need to be studied thoroughly to be able to offer corresponding and permanent services. The requested link to the incentive system demands the adaption of such frameworks and the metrics used. New metrics are already being explored and tested: altmetrics[202] for example, points to the initiative to incorporate more than articles, e.g. datasets, code, nanopublications[203]. These services need to be investigated more thoroughly in the future. Such approaches are needed to allow for a reconsideration of the current academic incentive and research assessment schemes.

[201] See for example the statements by the European Commission (2012a).

[202] http://altmetrics.org/manifesto/ [accessed August 19, 2012].

[203] http://nanopub.org/wordpress/ [accessed August 19, 2012].

Bibliography

Akopov, Z., Amerio, S., Asner, D., et al. (2012). Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics. Retrieved from http://arxiv.org/abs/1205.4667

Allianz der deutschen Wissenschaftsorganisationen (2010). Grundsätze zum Umgang mit Forschungsdaten. Retrieved from http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten/grundsaetze/

Alsheikh-Ali, A., Qureshi, W., Al-Mallah, M. H., Ioannidis, J. P. (2011). Public Availability of Published Research Data in High-Impact Journals. *PLOS ONE*, 6(9), e24357. doi:10.1371/journal.pone.0024357

Aymar, R. (2009). Scholarly Communication in High-Energy Physics: Past, Present and Future Innovations. *European Review*, 17(01), 33. doi:10.1017/S1062798709000556

Bell, S., Foster, N. F., & Gibbons, S. (2005). Reference librarians and the success of institutional repositories. *Reference Services Review*, 33(3), 283–290. doi:10.1108/00907320510611311

Birney, E., Hudson, T. J., Green, E. D., et al. (2009). Prepublication data sharing. *Nature*, 461(7261), 168–70. doi:10.1038/461168a

Björk, B.-C., Welling, P., Laakso, M., Majlender, P., Hedlund, T., & Guðnason, G. (2010). Open Access to the Scientific Journal Literature: Situation 2009. *PLOS ONE*, 5(6), e11273. doi:10.1371/journal.pone.0011273

Björk, B.-C., & Solomon, D. (2012). Open access versus subscription journals: a comparison of scientific impact. *BMC Medicine*, 10(1), 73. doi:10.1186/1741-7015-10-73

Bogner, A., Littig, B., & Menz, W. (2005). Das Experteninterview: Theorien, Methoden, Anwendungsfelder (2.Ed.) Wiesbaden, Germany. VS Verlag für Sozialwissenschaften. ISBN: 9783531144474

Borgman, C. L. (2010). Research Data: Who will share what, with whom, when, and why? China-North American Library Conference. Beijing, China. Retrieved from http://works.bepress.com/cgi/viewcontent.cgi?article=1237&context=borgman

Borgman, C. L. (2012). The Conundrum of Sharing Research Data. Journal of the American Society for Information Science and Technology, 63(6), 1059–1078. doi:10.1002/asi.22634

Brooks, T. C., Carli, S., Dallmeier-Tiessen, S., Mele, S., & Weiler, H. (2011). Authormagic in INSPIRE Author Disambiguation in Scholarly Communication. In *Proceedings of the ACM WebSci'11*, June 14-17 2011, Koblenz, Germany. (pp 1-2). Retrieved from http://journal.webscience.org/485/1/158_paper.pdf

Brumfiel, G. (2011). High Energy Physics: Down the petabyte highway. *Nature*, 469(7330), 282–283. doi:10.1038/469282a

Büttner, S., Hobohm, H., & Müller, L. (Eds.) (2011). Handbuch Forschungsdatenmanagement. Bad Honnef, Germany: Bock + Herchen. ISBN: 9783883472836. urn:nbn:de:kobv:525-opus-2412.

Campbell, E. G., Clarridge, B. R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N. A., & Blumenthal, D. (2002). Data withholding in academic genetics: evidence from a national survey. *JAMA: the journal of the American Medical Association*, 287(4), 473–480.

Carlson, D. (2011). A lesson in sharing. *Nature*, 469(7330), 293. doi:10.1038/469293a

Carpenter, J., Wetheridge, L., Tanner, S., & Smith, N. (2012). Researchers of Tomorrow: the research behaviour of Generation Y doctoral students Acknowledgements. Retrieved from http://www.jisc.ac.uk/media/documents/publications/reports/2012/researchers-of-tomorrow.pdf

Chavan, V., & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, 12(Suppl 15), S2. doi:10.1186/1471-2105-12-S15-S2

Collins, E., & Hide, B. (2010). Use and relevance of Web 2.0 resources for researchers. In *Proceedings of the 14th International Conference on Electronic Publishing* 16-18 June 2010, Helsinki, Finland. pp. 271 – 289. Retrieved from http://elpub.scix.net/cgi-bin/works/Show?_id=119_elpub2010

Committee for a Study on Promoting Access to Scientific and Technical Data for the Public Interest, National Research Council (1999). A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases. Washington (US): National Academy Press. ISBN: 9780309068253. Retrieved from http://books.nap.edu/catalog.php?record_id=9692

CMS collaboration (2012). CMS data preservation , re-use and open access policy (p. 4). Retrieved from https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/RetrieveFile?docid=6032&version=1&filename=CMSDataPolicy.pdf

Creaser, C., Fry, J., Greenwood, H., Oppenheim, C., Probets, S., Spezi, V., & White, S. (2010). Authors' Awareness and Attitudes Toward Open Access Repositories. *New Review of Academic Librarianship*, 16(sup1), 145–161. doi:10.1080/13614533.2010.518851

Credit where credit is overdue [Editorial]. (2009). *Nature Biotechnology*, 27(7), 579. doi:10.1038/nbt0709-579

Cullen, R., & Chawner, B. (2010). Institutional repositories: assessing their value to the academic community. *Performance Measurement and Metrics*, 11(2), 131–147. doi:10.1108/14678041011064052

Dallmeier-Tiessen, S., Darby, R., Görner, B. et al. (2010). Open Access Publishing - Models and Attributes. Retrieved from http://edoc.mpg.de/478647

Dallmeier-Tiessen, S. (2011). Strategien bei der Veröffentlichung von Forschungsdaten. In Büttner, S., Hobohm, H., & Müller, L. (Eds.). Handbuch Forschungsdatenmanagement (pp.

157–168). Bad Honnef, Germany: Bock + Herchen. ISBN: 9783883472836. urn:nbn:de:kobv:525-opus-2379.

Dallmeier-Tiessen, S. & Lengenfelder, A. (2011) Open Access in der deutschen Wissenschaft – Ergebnisse des EU-Projekts „Study of Open Access Publishing“ (SOAP), *GMS Med Bibl Inf,* 11(1-2), doi: 10.3205/mbi000218

Dallmeier-Tiessen, S., Darby, R., Görner, B., et al. (2011a). Dataset of the SOAP survey. Supplement to “Highlights from the SOAP project survey. What Scientists Think about Open Access Publishing”. Retrieved from http://arxiv.org/abs/1101.5260

Dallmeier-Tiessen, S., Darby, R., Görner, B., et al. (2011b). Highlights from the SOAP project survey. What Scientists Think about Open Access Publishing. Retrieved from http://arxiv.org/abs/1101.5260

Dallmeier-Tiessen, S., Darby, R., Görner, B., et al. (2011c). Open access journals–what publishers offer, what researchers want. *Information Services and Use*, 31(1), 85–91. doi:10.3233/ISU-2011-0624

Dallmeier-Tiessen, S., Pfeiffenberger, H., Schäfer, A., Pampel, H. (2011d). Ten Tales of Drivers & Barriers in Data Sharing. Retrieved from http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/10/7836_ODE_brochure_final.pdf

Dallmeier-Tiessen, S. (2012). Die wissenschaftsorientierte Publikation von Forschungsdaten. In Hohoff, U. & Lülfing, D. (Eds.). Bibliotheken für die Zukunft - Zukunft für die Bibliotheken. 100. Deutscher Bibliothekartag in Berlin 2011 (pp. 75–86). Hildesheim, Germany: Georg Olms Verlag.

Dallmeier-Tiessen, S., Darby, R., Gitmans, K., Lambert, S., Suhonen, J. A., & Wilson, M. (2012). Compilation of results on drivers and barriers and new opportunities. Retrieved from http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/08/ODE-CompilationResultsDriversBarriersNewOpportunities1.pdf

Data’s shameful neglect [Editorial].(2009). *Nature*, 461(7261), 145. doi:10.1038/461145a

Davis, P. M., & Connolly, M. J. L. (2007). Institutional Repositories: Evaluating the Reasons for Non-use of Cornell University’s Installation of DSpace. *D-Lib Magazine*, 13(3-4). Retrieved from http://www.dlib.org/dlib/march07/davis/03davis.html

Degkwitz, A.; Schirmbacher, P. (Eds.)(2007). Informationsinfrastrukturen im Wandel : Informationsmanagement an deutschen Universitäten. Bad Honnef: Bock & Herchen. ISBN: 9783883472546. Retrieved from http://www.dini.de/fileadmin/docs/DINI_Informationsinfrastrukturen.pdf

Deutsche Forschungsgemeinschaft (1998). Vorschläge zur Sicherung guter wissenschaftlicher Praxis. Retrieved from http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf

Deutsche Forschungsgemeinschaft (2009). Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten. Retrieved from http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf

Deutsche Forschungsgemeinschaft (2010). Leitfaden für die Antragsstellung: Projektanträge. Retrieved from http://www.dfg.de/formulare/54_01/54_01_de.pdf

European Commission (2012a). Commission Recommendation on access to and preservation of scientific information. C(2012) 4890 final. Retrieved from http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf

European Commission (2012b). Online survey on scientific information in the digital age. Luxembourg: Publications Office of the European Union. Retrieved from http://ec.europa.eu/research/science-society/document_library/pdf_06/survey-on-scientific-information-digital-age_en.pdf

Friend, S. H. (2010). Something in common. *Science translational medicine*, 2(40), 40ed6. doi:10.1126/scitranslmed.3001280

Fry, J., & Talja, S. (2007). The intellectual and social organization of academic fields and the shaping of digital resources. *Journal of Information Science*, 33(2), 115–133.

Fry, J., Probets, S., Creaser, C., Greenwood, H., Spezi, V., & White, S. (2011). Final Report: PEER Behavioural Research: Authors and Users vis-à-vis Journals and Repositories. Retrieved from http://www.peerproject.eu/fileadmin/media/reports/PEER_D4_final_report_29SEPT11.pdf

Gentil-Beccot, A., Mele, S., & Brooks, T. (2009). Citing and Reading Behaviours in High Energy Physics. How a Community Stopped Worrying about Journals and Learned to Love Repositories. Retrieved from http://arxiv.org/abs/0906.5418

Gentil-Beccot, A., Mele, S., Holtkamp, A., O'Connell, H. B., & Brooks, T. C. (2009). Information resources in High Energy Physics: Surveying the present landscape and charting the future course. *Journal of the American Society for Information Science and Technology*, 60(1), 150–160. doi:10.1002/asi.20944

Ginsparg, P. (2011). ArXiv at 20. *Nature*, 476(7359), 145–147. doi:10.1038/476145a

Goldschmidt-Clermont, L. (1965/2002). Communication Patterns in High-Energy Physics. *High Energy Physics Libraries Webzine*, (6). Retrieved from http://eprints.rclis.org/handle/10760/4253

Heath, T., & Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space (1.Ed.). San Francisco (US): Morgan & Claypool. doi:10.2200/S00334ED1V01Y201102WBE001

Heuer, R., Holtkamp, A., Mele, S. (2008). Innovation in scholarly communication: Vision and projects from High-Energy Physics. *Information Services and Use*, 28(2), 83-96, doi: 10.3233/ISU-2008-0570

Hodson, S. (2009). Data-sharing culture has changed. Research Information. Retrieved from http://www.researchinformation.info/features/feature.php?feature_id=243

Holley, R. (2010). Crowdsourcing: How and Why Should Libraries Do It? *DLib Magazine*, 16(3/4). doi:10.1045/september2004-vandesompel

Holzner, A., Igo-Kemenes, P., & Mele, S. (2009). First results from the PARSE.Insight project: HEP survey on data preservation, re-use and (open) access. Retrieved from http://arxiv.org/abs/0906.0485

Hrynaszkiewicz, I., & Cockerill, M. J. (2012). Open by default: a proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals. *BMC research notes*, 5(1), 494. doi:10.1186/1756-0500-5-494

Hurd, J. M. (2000). The transformation of scientific communication: A model for 2020. *Journal of the American Society for Information Science*, 51(14), 1279–1283. doi:10.1002/1097-4571(2000)9999:9999<::AID-ASI1044>3.0.CO;2-1

Igo-Kemenes, P. , Mele, S., Holzner, A. et al. (2010). Insight into digital preservation of research output in Europe: Case studies report (D3.3). Retrieved from http://www.parse-insight.eu/downloads/PARSE-Insight_D3-3_CaseStudiesReport.pdf

Ince, D. C., Hatton, L., & Graham-Cumming, J. (2012). The case for open computer programs. *Nature*, 482(7386), 485–488. doi:10.1038/nature10836

Key Perspectives (2010). Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long term Viability: SCARP Synthesis Study. Digital Curation Centre. Retrieved from http://www.era.lib.ed.ac.uk/bitstream/1842/3364/1/SCARP SYNTHESIS.pdf

Kroes, N. (2010). Unlocking the full value of scientific data. Speech/10/518. Retrieved from http://europa.eu/rapid/pressReleasesAction.do?reference=SPEECH/10/518&format=HTML&aged=0&language=EN&guiLanguage=en

Kuipers, T & van der Hoeven, J. (2009). Insight into digital preservation of research output in Europe: survey report (D3.4). Retrieved from http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

Kvenild, C., & Calkins, K. (2011). Embedded Librarians: Moving Beyond One-Shot Instruction (1 Ed.). Chicago, US: American Library Association. ISBN: 9780838985878

Laakso, M., Welling, P., Bukvova, H., Nyman, L., Björk, B.-C., & Hedlund, T. (2011). The development of open access journal publishing from 1993 to 2009. *PLOS ONE*, 6(6), e20961. doi:10.1371/journal.pone.0020961

Maron, N. L., & Smith, K. K. (2009). Current Models of Digital Scholarly Communication: Results of an Investigation Conducted by Ithaka Strategic Services for the Association of Research Libraries. *Journal of Electronic Publishing*, 12(1). Retrieved from http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0012.105

McVeigh, M. (2004). Open Access Journals in the ISI Citation Databases: Analysis of Impact Factors and Citation Patterns. Thomson Scientific. Retrieved from http://biblioteca.uned.es/lenya/bibliuned/search-authoring/docpdf/oacitations2.pdf

Meyer, E., Bulger, M., Zacharoudiou, A., Power, L., Williams, P., Venters, W., & Terras, M. (2011). Collaborative Yet Independent: Information Practices in the Physical Sciences. Research Information Network (RIN) Report Series, IOP Publishing. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1991753

Morris, S., & Thorn, S. (2009). Learned society members and open access. *Learned Publishing*, 22(3), 19. doi: http://dx.doi.org/10.1087/2009308

Mulligan, A., & Mabe, M. (2011). The effect of the internet on researcher motivations, behaviour and attitudes. *Journal of Documentation*, 67(2), 290–311. doi:10.1108/00220411111109485

Müller, U. (2009). Peer-Review-Verfahren zur Qualitätssicherung von Open-Access-Zeitschriften: Systematische Klassifikation und empirische Untersuchung. Dissertation, Humboldt-Universität zu Berlin. Retrieved from http://edoc.hu-berlin.de/docviews/abstract.php?lang=ger&id=29636. urn:nbn:de:kobv:11-10096430

Nelson, B. (2009). Data sharing: Empty archives. *Nature*, 461(7261), 160–163. doi:10.1038/461160a

Nentwich, M. (2003). Cyberscience – Research in the Age of the Internet. Vienna, Austria: Austrian Academy of Science Press. ISBN: 9783700131885, Retrieved from: http://hw.oeaw.ac.at/3188-7inhalt?frames=yes

Nentwich, M. (2009). Cyberscience 2.0 oder 1.2? Das Web 2.0 und die Zukunft der Wissenschaft. *ITA manu:scripts*, Institut für Technikfolgen-Abschätzung, ITA-09-02. Retrieved from: http://epub.oeaw.ac.at/ita/ita-manuscript/ita_09_02.pdf

Nicholas, D., Rowlands, I., Watkinson, A., Brown, D., & Jamali, H. R. (2012). Digital repositories ten years on: what do scientific researchers think of them and how do they use them? *Learned Publishing*, 25(3), 195–206. doi:10.1087/20120306

Pampel, H., & Bertelmann, R. (2011). „Data Policies“ im Spannungsfeld zwischen Empfehlung und Verpflichtung. In Büttner, S., Hobohm, H., & Müller, L. (Eds.)(2011). Handbuch Forschungsdatenmanagement. (pp. 49–61) Bad Honnef, Germany: Bock + Herchen. ISBN: 9783883472836. urn:nbn:de:kobv:525-opus-2287.

Pampel, H., Bertelmann, R., & Hobohm, H.-C. (2010). “Data Librarianship” - Rollen, Aufgaben, Kompetenzen. In Hohoff, U. & Schmiedeknecht, C. (Eds.). Ein neuer Blick auf Bibliotheken (pp. 159–176). Hildesheim, Germany: Olms. Retrieved from http://econpapers.repec.org/paper/rswrswwps/rswwps144.htm

Parse.Insight Consortium (2007). First insights into digital preservation of research output in Europe. Interim Insight Report. Retrieved from: http://www.parse-insight.eu/downloads/PARSE-Insight_D3-5_InterimInsightReport_final.pdf

Pelizzari, E. (2004). Academic authors and open archives: A survey in the social science field. *Libri*, 54, 113–122. Retrieved from http://www.librijournal.org/pdf/2004-2pp113-122.pdf

Pickard, A. J. (2007). Research methods in information. London, UK: Facet Publishing. ISBN: 9781856045452.

Piwowar, H. A. (2011a). Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. *PLOS ONE*, 6(7), e18657. doi:10.1371/journal.pone.0018657

Piwowar H.A., Day R.S., Fridsma D.B. (2007) Sharing detailed research data is associated with increased citation rate. *PLOS ONE,* 2(3), e308. doi: 10.1371/journal.pone.0000308

Procter, R., & Williams, R. (2010). If You Build It, Will They Come? How Researchers Perceive and Use Web 2.0. Research Information Network. Retrieved from: http://rinarchive.jisc-collections.ac.uk/our-work/communicating-and-disseminating-research/use-and-relevance-web-20-researchers

Procter, R., Williams, R., Stewart, J., Poschen, M., Snee, H., Voss, A., & Asgari-Targhi, M. (2010). Adoption and use of Web 2.0 in scholarly communications. *Philosophical transactions A. Mathematical, physical, and engineering sciences,* 368(1926), 4039–4056. doi:10.1098/rsta.2010.0155

Pryor, G. (2012). Managing Research Data. London, UK: Facet Publishing. ISBN: 9781856047562

Roosendaal, H. E., & Geurts, P. A. T. M. (1997). Forces and functions in scientific communication: an analysis of their interplay. *CRISP 97 Cooperative Research Information Systems in Physics*, d(April), 1–32. Retrieved from http://doc.utwente.nl/60395/

Rowlands, I., & Nicholas, D. (2005). Scholarly communication in the digital environment: The 2005 survey of journal author behaviour and attitudes. *ASLIB Proceedings*, 57(6), 481–497. doi:10.1108/00012530510634226

De la Sablonnière, R. D., Auger, E., Sabourin, M., & Newton, G. (2012). Facilitating data sharing in the behavioural sciences. *Data Science Journal*, 11. 1–15. doi:http://dx.doi.org/10.2481/dsj.11-DS4

Sansone, S.-A., Rocca-Serra, P., Field, D., et al. (2012). Toward interoperable bioscience data. *Nature* Genetics, 44(2), 121–126. doi:10.1038/ng.1054

Savage, C. J. & Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. *PLOS ONE*, 4(9), e7078. doi:10.1371/journal.pone.0007078

Schofield, P. N., Bubela, T., Weaver, T., et al. (2009). Post-publication sharing of data and tools. *Nature*, 461(7261), 171–173. doi:10.1038/461171a

Schäfer, A., Pampel, H., Pfeiffenberger, H. et al. (2011). Baseline Report on Drivers and Barriers in Data Sharing. Retrieved from http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-WP3-DEL-0002-1_0_public_final.pdf

Science. (2011). Special Online Collection: Dealing with Data. Retrieved from http://www.sciencemag.org/site/special/data/

Shieber, S. (2009). What percentage of open-access journals charge publication fees? The Occasional Pamphlet. Retrieved from http://blogs.law.harvard.edu/pamphlet/2009/05/29/what-percentage-of-open-access-journals-charge-publication-fees/

Smith, D. & Carrano, A. (1996). International Large-Scale Sequencing Meeting. *Human Genome News*, 6(7). Retrieved from http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v7n6/19intern.shtml

Spier, R. (2002). The history of the peer-review process. *Trends in biotechnology*, 20(8), 357–358.

Study Group for Data Preservation and Long Term Analysis in High Energy Physics, DPHEP (2009). Data Preservation in High Energy Physics. Retrieved from http://arxiv.org/abs/1101.3186

Suber, P. & Sutton, C. (2007). Society publishers with open access journals. Retrieved August 2, 2012, from http://www.earlham.edu/~peters/fos/newsletter/11-02-07.htm#list

Suber, P. (2012). Open Acces. Cambridge, US: MIT Press. ISBN: 9780262517638

Swan A. & Brown, S. (2004). JISC/OSI JOURNAL AUTHORS SURVEY. Retrieved from http://www.jisc.ac.uk/uploaded_documents/ACF655.pdf

Swan, A. & Brown, S. (2008). To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. Report commissioned by the Research Information Network (RIN). Retrieved from http://eprints.soton.ac.uk/id/eprint/266742

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., et al. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLOS ONE*, 6(6), e21101. doi:10.1371/journal.pone.0021101

Thomson Reuters (2012). Thomson Reuters Unveils Data Citation Index for Discovering Global Data Sets | Reuters. Retrieved from http://www.reuters.com/article/2012/06/22/idUS109861+22-Jun-2012+HUG20120622

Van Berchum, M. (2011). The results of the SOAP survey. A first overview of the Duch situation. Retrieved from http://www.openaccess.nl/images/pdf/soap_nl.pdf

Van De Sompel, H., Payette, S., Erickson, J., Lagoze, C., & Warner, S. (2004). Rethinking Scholarly Communication: Building the System that Scholars Deserve. *DLib Magazine*, 10(9). doi:10.1045/september2004-vandesompel

Van Noorden, R. (2012). Journal offers flat fee for “all you can publish”. *Nature*, 486(7402), 166. doi:10.1038/486166a

Walshe, K. (2011). Defining a new role: the embedded Research Information Manager. Retrieved from http://www.jisc.ac.uk/media/documents/programmes/RIM/RIMReport_FINAL.pdf

Ware, M. (2008). Peer review: benefits, perceptions and alternatives. Publishing Research Consortium. Retrieved from http://www.publishingresearch.net/documents/PRCsummary4Warefinal.pdf

Ware, M. & Mabe, M. (2009). The STM report. An overview of scientific and scholarly journal journal publishing. Retrieved from http://www.stm-assoc.org/2009_10_13_MWC_STM_Report.pdf

Warlick, S. E., & Vaughan, K. T. L. (2007). Factors influencing publication choice: why faculty choose open access. *Biomedical digital libraries*, 4(1), 1. doi:10.1186/1742-5581-4-1

Weiler, H. (2012). Authormagic: A Concept for Author Disambiguation in Large-Scale Digital Libraries. Dissertation. University of Erlangen-Nuremberg. Retrieved from http://www.opus.ub.uni-erlangen.de/opus/volltexte/2012/3139/

Wellcome Trust. (2003). Sharing data from large-scale biological research projects: a system of tripartite responsibility. Fort Lauderdale. Retrieved from http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf

Whyte, A. & Rusbridge, C. (2010). Digital Curation Centre User Survey 2009: Highlights. Retrieved from http://digitalcuration.blogspot.com/2010/01/digital-curation-centre-user-survey.html

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. *PLOS ONE*, 6(11), e26828. doi:10.1371/journal.pone.0026828

Wickham, J. (2011). Unlocking attitudes to Open Access – survey results | Repositories Support Project on WordPress.com. Retrieved from http://rsppproject.wordpress.com/2011/12/02/unlocking-attitudes-to-open-access-survey-results/

Williams, P., Stevenson, L., Nicholas, D., Watkinson, A., & Rowlands, I. (2009). The role and future of the monograph in arts and humanities research. *ASLIB Proceedings*, 61(1), 67–82.

Wolkovich, E. M., Regetz, J., & O'Connor, M. I. (2012). Advances in global change research require open science by individual researchers. *Global Change Biology*, 18(7), 2102–2110.

Referenced Interviews (Chapter 3)

Participant 1 (2011), conducted by the author of this thesis. Geneva. 03.03.2011

Participant 2 (2011), conducted by the author of this thesis. Geneva. 07.03.2011

Participant 3 (2011), conducted by the author of this thesis. Geneva. 11.03.2012

Participant 4 (2011), conducted by the author of this thesis. Geneva. 16.03.2012

Participant 5 (2011), conducted by the author of this thesis. Geneva. 18.03.2011

Participant 6 (2011), conducted by the author of this thesis. Geneva. 07.04.2011

Participant 7 (2011), conducted by the author of this thesis. Geneva. 14.04.2011

Participant 8 (2012), conducted by the author of this thesis. Geneva. 21.02.2012

Participant 9 (2012), conducted by the author of this thesis. Geneva. 21.02.2012

Participant 10 (2012), conducted by the author of this thesis. Geneva. 21.02.2012

Participant 11 (2012), conducted by the author of this thesis. Geneva. 22.02.2012

Participant 12 (2012), conducted by the author of this thesis. Geneva. 23.02.2012

Participant 13 (2012), conducted by the author of this thesis. Geneva. 23.02.2012

Participant 14 (2012), conducted by the author of this thesis. Geneva. 23.02.2012

Participant 15 (2012), conducted by the author of this thesis. Geneva. 24.02.2012

Participant 16 (2012), conducted by the author of this thesis. Geneva. 24.02.2012

Participant 17 (2012), conducted by the author of this thesis. Geneva. 02.03.2012

Participant 18 (2012), conducted by the author of this thesis. Geneva. 02.03.2012

Participant19 (2012), conducted by the author of this thesis. Geneva. 05.03.2012

Participant 20 (2012), conducted by the author of this thesis. Geneva.06.03.2012

Participant 21 (2012), conducted by the author of this thesis. Geneva. 08.03.2012

Participant 22 (2012), conducted by the author of this thesis. Geneva. 09.03.2012

Participant 23 (2012), conducted by the author of this thesis. Geneva. 20.03.2012

Participant 24 (2012), conducted by the author of this thesis. Geneva. 03.04.2012

Participant 25 (2012), conducted by the author of this thesis. Geneva. 06.04.2012

Participant 26 (2012), conducted by the author of this thesis. Geneva. 11.05.2012

List of Figures

List of Tables

Appendix A: Supplementary Materials to Chapter 2

The dataset of the survey has been published as Dallmeier-Tiessen et al. (2011a). Alongside the full survey including all questions and possible answers has been shared. The definition of gold Open Access has been stated at the beginning of the survey: "For the purposes of this survey, an article is Open Access if its final, peer-reviewed, version is published online by a journal and is free of charge to all users without restrictions on access or use." The dataset has been released with a Creative Commons Zero waiver (public domain dedication). The release is documented with a detailed data manual available as supplemental material to the article.

Table 1: Representation of researchers by discipline.

Representation of disciplines in the SOAP survey. Respondents chose their domain via a drop down menu. Only published researchers (at least one peer reviewed article) with a response to question 9 (Table 2) have been included in the analysis presented in this thesis (n=37100).

Disciplines	**Number of respondents**
Medicine, Dentistry and Related Subjects	7010
Biological Sciences	6942
Social Sciences	3296
Mathematical and Computer Sciences	3189
Physics and Related Sciences	2642
Engineering and Technology	2518
Chemistry	1723
Psychology	1602
Earth Sciences	1434
Education	1358
Agriculture and Related Sciences	1109
Business and Administrative Studies	1066
Historical and Philosophical Studies	932
Language and Literature Studies	675
Astronomy and Space Science	587
Mass Communications and Documentation	469
Architecture, Building and Planning	229
Law	197
Creative Arts and Design	121

Table 2: Response to question 9 to the survey.

The numbers given comprise only published researchers, who completed question 9 (n=37100): "Do you think your research field benefits, or would benefit from journals that publish Open Access articles?"

Discipline	Yes	No opinion/I don't care	No
Language and Literature Studies	95.3%	3.6%	1.2%
Creative Arts and Design	94.2%	5.0%	0.8%
Mass Communications and Documentation	94.0%	4.1%	1.9%
Education	93.0%	5.1%	1.9%
Historical and Philosophical Studies	92.0%	4.9%	3.1%
Social Sciences	91.9%	5.9%	2.2%
Medicine, Dentistry and Related Subjects	91.7%	5.7%	2.7%
Biological Sciences	91.6%	5.2%	3.2%
Psychology	90.0%	6.5%	3.5%
Law	89.8%	7.6%	2.5%
Business and Administrative Studies	89.6%	7.3%	3.1%
Earth Sciences	89.5%	7.2%	3.3%
Agriculture and Related Sciences	88.9%	9.4%	1.7%
Mathematical and Computer Sciences	87.0%	7.9%	5.0%
Architecture, Building and Planning	86.9%	11.4%	1.7%
Engineering and Technology	84.8%	10.4%	4.7%
Astronomy and Space Science	83.0%	9.2%	7.8%
Physics and Related Sciences	82.7%	11.4%	5.9%
Chemistry	76.9%	14.3%	8.8%

Table3: Response to question 13 of the survey.

Response to the question: "What factors are important to you when selecting a journal to publish in?" Rating from Extremely Important to Irrelevant (respondents per factor vary, see Dallmeier-Tiessen et al. (2011a).

Factor	Extremely important	Important	Less important	Irrelevant
The journal fits the policy of my organisation	8.0%	27.4%	33.8%	30.8%
Copyright policy of the journal	8.0%	27.4%	43.8%	20.7%
The journal is Open Access	10.7%	33.5%	39.4%	16.3%
Recommendation of the journal by my colleagues	10.9%	45.8%	34.1%	9.1%
Absence of journal publication fees (e.g. submission charges, page charges, colour charges)	29.2%	37.3%	26.4%	7.1%
Importance of the journal for academic promotion, tenure or assessment	32.3%	43.1%	18.2%	6.4%
Positive experience with publisher/editor(s) of the journal	22.5%	56.0%	18.1%	3.5%
Speed of publication of the journal	24.9%	53.6%	19.2%	2.3%
Likelihood of article acceptance in the journal	20.9%	58.0%	18.2%	2.9%
Journal Impact Factor	36.9%	47.2%	13.0%	2.9%
Relevance of the journal for my community	50.6%	40.0%	7.3%	2.1%
Prestige/perceived quality of the journal	49.5%	44.4%	5.4%	0.7%

Table 4: Response to question 15 of the survey.

Responses to the question: "Approximately how many Open Access articles have you published in the last five years?" (only published researchers, who completed question 9 and 15 are taken into account, n=36334).

Discipline	0	1 to 5	6 to 10	More than 10	I do not know	Grand Total
Agriculture and Related Sciences	26.4%	55.5%	7.9%	3.7%	6.5%	100.0%
Architecture, Building and Planning	37.7%	44.5%	4.1%	3.2%	10.5%	100.0%
Astronomy and Space Science	41.3%	31.5%	5.2%	3.3%	18.6%	100.0%
Biological Sciences	19.3%	64.5%	8.3%	3.2%	4.7%	100.0%
Business and Administrative Studies	38.9%	39.4%	3.5%	3.4%	14.8%	100.0%
Chemistry	41.5%	40.1%	4.6%	4.9%	8.9%	100.0%
Creative Arts and Design	20.8%	60.0%	5.0%	4.2%	10.0%	100.0%
Earth Sciences	28.9%	56.0%	5.5%	3.4%	6.2%	100.0%
Education	29.8%	50.6%	3.7%	4.2%	11.7%	100.0%
Engineering and Technology	42.2%	40.9%	4.5%	2.2%	10.2%	100.0%
Historical and Philosophical Studies	35.7%	47.7%	4.6%	1.6%	10.4%	100.0%
Language and Literature Studies	31.4%	51.2%	5.1%	1.5%	10.7%	100.0%
Law	27.2%	55.0%	3.1%	5.2%	9.4%	100.0%
Mass Communications and Documentation	26.6%	58.6%	5.7%	2.0%	7.2%	100.0%
Mathematical and Computer Sciences	33.7%	45.9%	4.9%	2.7%	12.8%	100.0%
Medicine, Dentistry and Related Subjects	19.2%	61.7%	8.5%	3.9%	6.8%	100.0%
Physics and Related Sciences	32.5%	42.3%	7.3%	5.2%	12.7%	100.0%
Psychology	38.9%	41.8%	4.2%	2.0%	13.2%	100.0%
Social Sciences	33.5%	49.2%	4.1%	1.7%	11.5%	100.0%

Table 5: Response to question 16 of the survey.

Free text answers to question 16: "Has there been a specific reason not to publish OA?". The free text answers have been analyzed, tagged and grouped. Only disciplines with more than 100 tags are shown (n=5609 tags).

Discipline	Access-ibility	Journal quality	Fun-ding	Next time	Habits	Other	Un-awareness	Total
Biological Sciences	4%	20%	59%	2%	4%	7%	3%	100%
Agriculture and Related Sciences	5%	22%	54%	0%	5%	8%	5%	100%
Medicine, Dentistry and Related Subjects	6%	23%	53%	2%	4%	7%	5%	100%
Chemistry	3%	39%	44%	1%	2%	6%	4%	100%
Mathematical and Computer Sciences	8%	31%	42%	1%	3%	11%	4%	100%
Physics and Related Sciences	7%	29%	40%	2%	5%	12%	5%	100%
Engineering and Technology	8%	35%	39%	2%	3%	5%	7%	100%
Earth Sciences	9%	32%	38%	4%	5%	7%	5%	100%
Psychology	11%	30%	28%	4%	5%	11%	13%	100%
Historical and Philosophical Studies	14%	26%	27%	5%	7%	13%	8%	100%
Education	11%	26%	23%	4%	5%	15%	17%	100%
Social Sciences	11%	34%	22%	3%	5%	15%	12%	100%
Astronomy and Space Science	10%	49%	16%	1%	6%	14%	5%	100%
Business and Administrative Studies	12%	37%	12%	4%	4%	15%	16%	100%

Appendix B: Supplementary Materials to Chapter 3

First round of interviews

Interview guideline for first round of interview (see also details in project report: Schäfer et al., 2011):

In preparation for the interview following information about the interviewee should be recorded:

1. brief introduction of the interview partner (organization, position, role)

2. nature of research data in this person's sphere

3. perceived state of dealing with research data in this person's sphere

In the interview the following aspects should be addressed:

- highlights in data sharing
- lowlights in data sharing
- unforeseen events in data sharing
- intentions for the future sharing of data

During the interview the interviewer should asked for factors, which influenced the highlights or lowlights. The factors may be of financial, technical, legal and social nature.

Based on this general rule, individual questions for the interviewers were developed based on their expertise and research domain. The complete results of the first round of interviews are given in the report of the project (see Dallmeier-Tiessen et al., 2011d and Schäfer et al., 2011). Again it needs to be emphasized that this round of interview aimed at recording the individual experience with research data sharing.

Workshop

The workshop took place at BMA House in London, November 7th, 2011. It was held alongside the Alliance for Permanent Access conference which took place in the same venue that week. Invited participants had complementary expertise in data related activities in different disciplines; they were selected by their publications or affiliation to expert groups. Furthermore they covered disciplines that were underrepresented in the first round of interviews (e.g. clinical trials, biodiversity science).

The participants were provided with the drivers and barriers identified in the first round of the interviews. The guided discussion was used to revise the list of drivers and barriers and to prepare the interview guideline for the second round of the interviews.

Second round of interviews

Interview guideline for second round of interview is enclosed below:

The guideline for interviews with non-researchers follows this structure as well, but addressing the interviewees as the ones working with researchers.

The questionnaire asked if they/or the researchers they work with share their data or not. The questions ask for the expert's opinion on the drivers and barriers identified in the first round and his experience in that regard. This means that the list of drivers and barriers (see chapter 3) was sent to him before the interview. In addition, examples of overcoming barriers are investigated in the second round of the interviews.

Name of interviewer	
Date of interview	

Method of interview (telephone, Skype, …)	
Approximate duration of interview (minutes)	

Questions to be filled in prior to interview from reply to invitation

1. Name	
2. Organisation	
3. Country	
4. Age group	
5. Role	
6. Academic discipline	

Multiple choice questions

7. Which of the following applies to the digital research data of your research: (multiple answers possible)?	a) My data is openly available for everyone. b) My data is available for a fee. c) My data is openly available for my research discipline. d) My data is openly available for my research group / colleagues in research collaboration. e) Access to my data is temporarily restricted Which restriction?
8. Which of the ODE drivers motivate you in sharing your data with others? (multiple answers possible from list)	a) Societal benefits b) Academic Benefits c) Research Benefits d) Organisational Incentives e) Individual Contributor Incentives
9. Which of the ODE drivers motivate you in using other people's data? (multiple answers possible from list)	a) Societal benefits b) Academic Benefits c) Research Benefits d) Organisational Incentives e) Individual Contributor Incentives
10. Do you presently/or in the past make use of research data gathered by other researchers within	Yes / No

your discipline?	
11. Which of the ODE barriers have you encountered in sharing data within your discipline? (multiple answers possible from list)	f) Individual Contributor barriers g) Availability of a Sustainable Preservation Infrastructure h) Trustworthiness of the data, Data Usability, Pre-archive activities i) Data Discovery j) Academic Defensiveness k) Finance l) Subject Anonymity and Personal Data Confidentiality m) Legislation/Regulation
12. Do you presently/or in the past make use of research data gathered by other researchers in other disciplines?	Yes / No
13. Which of the ODE barriers have you encountered in sharing data outside your discipline? (multiple answers possible from list)	f) Individual Contributor barriers g) Availability of a Sustainable Preservation Infrastructure h) Trustworthiness of the data, Data Usability, Pre-archive activities i) Data Discovery j) Academic Defensiveness k) Finance

	l) Subject Anonymity and Personal Data Confidentiality m) Legislation/Regulation

Questions for free discussion

14. Can you give any examples of barriers preventing sharing, or of overcoming those barriers, in your discipline or across disciplines please? (one or more story answer(s)) *If the interviewee needs a prompt, then: what was the data, who was the data provider & data consumer, when did this happen, what was the research topic, why was the sharing important, what were the barriers, how were they overcome?*

15. From your perspective, what does good data citation look like, and why? *If the interviewee needs a prompt, then: This could be in terms of HOW (form of citation, identifiers used or technical solutions, WHERE (in reference lists, acknowledgements or in-line) and WHEN data is cited within a paper, or data citing resulting research.*

.

Appendix C: Supplementary Materials to Chapter 5

The numbers of crowdsourcing and engagement experiment are given in the following tables.

More details on the computing of the author disambiguation algorithm that is used for the crowdsourcing experiment can be found in Weiler (2012).

Table 6: Claimed artifacts in the respective time period (total numbers).

Response week (2011)	Claimed Artifacts (total)
11	444
12	1934
13	6594
14	9570
15	13437
16	16269
17	18945
18	21758
19	23276
20	25526
21	27495
22	29882
23	41582
24	44532
25	46619
26	48397
27	50058
28	51804
29	53260
30	61904

Response week (2011)	Claimed Artifacts (total)
31	63722
32	66154
33	67691
34	70940
35	71848
36	73132
37	103705
38	107584
39	110315
40	111911
41	132945
42	135552
43	147424
44	149339
45	150880
46	159896
47	164811

Table 7: Number of researchers who used the workflow in the respective time period in 2011 [in total 2558]. We can distinguish the once who have been invited versus the ones who found it by serendipitous discovery.

Week 2011	active researchers cumulative (who claimed artifacts, total count)	serendipitous discovery (total count)
11	5	5
12	23	23
13	73	73
14	101	101
15	145	145
16	176	176
17	205	205
18	233	233
19	253	253
20	277	277
21	297	297
22	321	321
23	435	361
24	465	387
25	485	406
26	503	422
27	519	437
28	538	456

Week 2011	active researchers cumulative (who claimed artifacts, total count)	serendipitous discovery (total count)
28	538	456
29	554	472
30	727	532
31	761	561
32	813	609
33	837	631
34	878	670
35	907	695
36	940	727
37	1440	884
38	1503	924
39	1549	951
40	1591	982
41	1864	1068
42	1912	1095
43	2080	1136
44	2126	1171
45	2161	1202
46	2505	1308
47	2558	1341

Table 8: Response to mail-out in hours after invitation [n=1,021]. The hourly timing of the mailing was not recorded for the first mail-outs. Thus the total number is lower than the numbers given above in table 7.

Time after mailout (hours)	People who claimed (total count)
0	306
1	156
2	50
3	50
4	58
5	32
6	14
7	20
8	9
9	11
10	5
11	6
12	4
13	2
14	2
16	3
17	8
18	2
19	13
20	8
21	11
22	18
23	19
24	12
25	7
26	10
27	7
28	3
29	2
30	3
31	3
32	4
33	1
34	2

Time after mailout (hours)	People who claimed (total count)
35	2
36	1
39	2
40	1
41	1
42	3
43	5
44	3
45	4
46	1
47	4
48	1
49	8
50	4
51	2
52	5
54	2
55	1
56	1
60	1
62	1
66	1
67	1
68	4
69	3
70	1
71	2
72	3
73	1
74	4

Time after mailout (hours)	People who claimed (total count)
76	2
78	1
80	2
81	1
82	1
83	2
86	1
89	1
90	1
93	1
97	2
98	1
100	1
101	2
102	1
104	2
112	1
113	1
116	3
119	2
120	1
121	3
123	1
126	1
127	2
128	1
136	1
138	1
140	1
142	1
143	1
144	2
145	1
146	2

Time after mailout (hours)	People who claimed (total count)
147	1
148	3
151	1
152	2
155	1
156	1
159	1
161	1
163	1
164	1
166	1
169	1
170	1
176	1
187	2
190	1
193	2
203	1
207	1
211	2
212	1
215	3
217	2
220	1
225	1
227	1
238	1
239	1
245	1
248	1

Declaration/Selbstständigkeitserklärung

Herewith, I declare on oath that this thesis is my independent achievement. It is based solely on the referenced sources and materials. Neither this thesis, nor a similar work has been submitted or published as a dissertation elsewhere.

Hiermit erkläre ich an Eides Statt, dass ich die vorliegende Arbeit selbständig verfasst und ausschliesslich unter Zuhilfenahme der angegebenen Quellen und Hilfsmittel verfasst habe. Ich versichere darüber hinaus, dass diese Arbeit in dieser oder eine anderen Form noch nicht anderweitig als Dissertation eingereicht oder veröffentlicht wurde.